Queueing Models With Finite and Infinite Buffering Capacity A Comparative Study.

by Hermin Anggawijaya
Supervised by : Dr. K. Pawlikowski

Abstract

Queueing theory offers a large variety of techniques that can be used in performance modelling of computer systems and data communication networks. The diversity of assumptions causes that the numerical results, obtained for the same system but by means of different techniques, can often be numerically very different. The project is aimed at finding common denominators for numerical results obtained with the assumption of infinite buffer capacity and that obtained with the assumption of finite buffer capacity.
To be precise, this project investigates the traffic intensity regions where the approximation of queueing systems with finite buffer capacity by the queueing systems with infinite buffer capacity can be done with some amount of safety margin.
The investigation includes both queueing systems with single arrivals (M/M/1, M/D/1), and a queueing system with batched arrivals ($M^{(b)}/M/1$), and two the most important characteristics of queueing systems are considered, the probability of overflow and average waiting time in the system.

Acknowledgements

I wish to thank Dr. K. Pawlikowski for providing help and guidance throughout the whole project. And also I wish to thank Dr. W. Kennedy for providing extra supervision while Dr. Pawlikowski was away.
Finally, I wish to thank my family for constant support and prayers, also my classmates, who always provide their helping hands.

Table of Contents

Chapter One

Introduction.

Queueing theory is a powerful tool for modelling computer systems and networks, and that can be used to obtain such characteristics as average response time, average number of jobs waiting for processing by the CPU, and so on.

Queueing theory has found applications in Computer Science due to the fact that, many phenomena in computer systems are characterised by the lack of certainty. For example, when a job is submitted to the system, generally we won't be sure how long is the time we need to wait before the job is completed, or the next job is submitted, or many jobs will be submitted during the next arrival, etc. We can model and analyse such uncertain processes by means of techniques offered by by queueing theory.

There are three basics elements that compose a queueing system. They are: *customer population*, where a customer can mean a print job, a packet to be transmitted over a communication line, a car waiting for refuelling at a gas station, or simply a person waiting to be served at a restaurant. Customers arrive in accordance with an arrival process modelled by, for example, Poisson probability distribution.
Server that provides service to customers. A server can be a CPU, a communication line, gas station attendant or a waiter in a restaurant. A server serves customers in a particular pattern. Customers can be served in first-in-first-served, or first-in-last-served, or randomly-served basis. If the server is busy, new customer(s) has to wait in a *waiting line* (a queue) which is the third basic element of any queueing system. Below is a picture of a very simple queueing system :

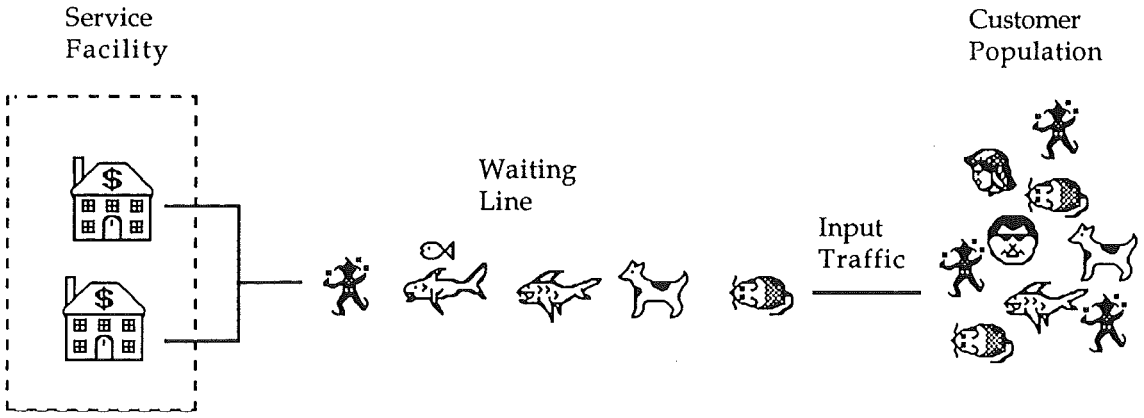


Fig . 1 A simple Queueing System

Population or Source of Customers

The population or source of customer in any queueing system is characterised by the size of it. Some systems are considered to cooperate with a finite population of customers, and others with infinitely large customer population. We need to clarify that the word *infinite* has a weaker sense, in that any customer population or source must be finite, but sometimes it is necessary to consider a very large population of customers as an infinite population. This sometimes turns out to be important factor in making the mathematical analysis of queueing system tractable. Finite populations - in modelling internal processes of computer systems. Infinite populations eg. in modelling data communication networks.

Customer Arrivals Pattern

The behaviour of queueing systems depends not only on the size of the population and the rates at which they arrive to the system but also on the pattern that they arrive. Generally, evenly spaced customer arrivals result in better performance of the system, because the service facility can provide a better service. The worst is when customers arrive to the system in clusters, the extreme case of clustered stream of customer arrivals is sometimes referred to as *batch arrival* pattern.

This project investigates both systems with single customer arrivals and systems with customers arriving in batches. However only exponentially distributed interarrival times are considered here. By exponentially distributed interarrival times we mean that the cumulative distribution function of customer interarrival times is

$$A[t] = 1 - e^{-\lambda t} \qquad (1)$$

where λ the average arrival rate. This arrival pattern possesses the so called *memoryless property*. If an arrival pattern possesses the memory less property, the next arrival time is completely independent of the present or past arrival times, see Allen [4]. Above, we have the symbol λ as the mean rate of customer arrivals, so it is obvious that $1/\lambda$ is the average interarrival time.

Service Time Distribution.

As mentioned above, this project investigates queueing systems with exponentially and deterministic service times distributions. For the exponential distribution of service times, the memoryless property means that the time remaining to complete the service of a customer is independent of the time already spent in servicing this particular customer. The service time distribution function is given by

$$W_s[t] = 1 - e^{-\mu t} = 1 - e^{\frac{-t}{W_s}} \qquad (2)$$

where $\mu = 1/Ws$, is the average rate at which a server processes customers when the server is busy
The other service time pattern considered here is deterministic service pattern, where customers require the same service time. The reason why that this service pattern is chosen, is that, it is very common situation in modern data communication systems ie. in transmission of standardised data blocks (packets of equal size).

Other Relevant Terminologies

Utilisation Coefficient is the term generally used for describing the degree of utilization of the service facility. It is defined as the proportion of average arrival rate to the average service rate, and the symbol that is used for utilisation factor is ρ. The behaviour of a queueing system is heavily dependent on the value of utilisation coefficient: the higher the utilisation factor the longer the average delay. Furthermore the level of utilisation affects the probability of buffer overflow.

Average System Delay . The average time the customer spends in the system, ie. from the arrival time of the customer to the its departure, and is defined as (from Little's formula [3]), for the exact formula see appendix A (A-1)

$$E[W] = \frac{\text{Mean number of customers in the system}}{\lambda} \qquad (3)$$

Average Queue Delay, denoted by the symbol W_q, is the average time that customer spent in the queue buffer, thus (for details see appendix A-2)

$$E[W_q] = \frac{\text{Mean queue length}}{\lambda} \qquad (4)$$

Note that the number of customer in the queue is equal to the number of customer in the system minus one, which is the average number of customer in the buffer queue (for single server queueing system). From there, using the same reasoning used in Little's formula we can deduce another relation namely : the total average time in the system equals to the average waiting time in queue plus the average service time, which is intuitively clear, ie.

$$E[W] = E[W_q] + E[W_s] \qquad (5)$$

This project investigates queueing systems with an assumed infinitely large customer population with poisson arrival pattern, both single and batch arrival case are studied.

Throughout this report we will refer to those queueing systems mentioned above with special kind of notation called Kendall notation, after David Kendall [1], its originator, specially developed to described queueing systems.
Kendall notation has the form of A/B/c/K/m/Z, where

A	describes the interarrival times distribution of customers
B	describes the service time distribution provided by server(s)
c	describes the number of servers in the system
K	describes the system buffer capacity
m	describes the number in the customer population
Z	describes the queue discipline used by the system

The symbols used for A and B in this project are, M stands for exponential interarrival or service time distribution with single arrival or service, $M^{(b)}$ stands for exponential interarrival time or service time distribution with batch arrivals or service, D stands for deterministic (constant) interarrival or service time distribution.

Chapter Two

The Queue Buffer Capacity.

Basic Assumption
Behaviour of any queueing system depends on the assumed size of the queue buffer in front of the server. Basically there are two options that can be considered : 1) Infinite buffer capacity and 2) Finite buffer capacity.

1) *Infinite Buffer Capacity.*
Some may argue that there is no buffer of infinite capacity. But the advantage of regarding a buffer as having infinite capacity is the simplicity of analysis. This effect is a consequence of the 'infinite sumability' of many mathematical models of such queueing systems. In many cases, the analysis can result in direct analytical formulas, we need to substitute only numbers to the formulas and get the results by performing simple mathematical computations.
Of course the results produced using such assumption, will only approximate the behaviour of finite queueing systems. But as will be shown latter, in some particular cases, the approximations are fairly accurate, eg. for low utilisation coefficient and large buffer capacities, the numerical approximations are fairly accurate.
A disadvantage of analysing systems with infinite buffers is that, such systems are stable only if they are utilised in less than (100%) thus for coefficient of utilisation $\rho < 1$. The reason for this, is that if we have unlimited buffer capacity and customer arrival rate is higher than the service rate provided by the service facility, then the queue system's buffer capacity will keep growing to an infinitely large size and the system become unstable.
In this project, we analyse the limiting behaviour of stable queueing systems, it is, their behaviour in steady state.

2) *Finite Buffer Capacity.*
One of the properties of finite buffer capacity queueing systems is that, most of them are mathematically difficult to analyse, since analysis leads to a set of equations, that has to be solved to get numerical solution . As the author experienced himself, solving equations for an infinite buffer queueing system is much more time consuming than making straight computation with the analytical formulas obtained from mathematical analysis of the queueing systems with unlimited buffer capacity.

On the other hand, with the analysis of such systems we can obtain exact numerical results with the accuracy being only limited by the precision of the computer used.

What to be Compared

Having considered all the main advantages and disadvantages of both assumptions about the queue system's buffer capacity, we have a question to ask, what happens if we use numerical results obtained for buffer of infinite capacity to analyse buffers with a finite capacity. That question summarised the main topic of this project.

To answer that question, first thing that we do is making comparisons on the performance measures of both systems with a certain range of system's buffer capacitys and utilisation.

We compare :

- The probabilities of the system being in a particular states
- Average numbers of customers in the systems
- Average of system delays (mean waiting times)
- The accuracy of approximations used for assessing the buffer overflow probability using numerical results taken from the infinite capacity buffers.

This project focuses on the last two comparisons, the first two comparisons are somewhat implied by the last two. If we have high probability of overflow than it is obvious that we can expect that the probability of large number of customer present in the system must be high. Similarly, the average number of customers in the system and the average system waiting time, this relationship is depicted in equation (3) in the Introduction chapter.

Method of Comparisons

Symbols related to buffers of finite capacity K will be distinguished from corresponding symbols related to buffers of infinite capacity by subscripts K and ∞, respectively. Thus W_K means the delay (the total time spent in queue) in a buffer of capacity K, while W_∞ means the delay on a buffer of infinite capacity.

Exact values of mean delays and overflow probabilities (obtained from analysis of finite buffers) will be distinguished from their approximations (based on results obtained from buffers of unlimited capacity) by means of dashing the letter parameter. Thus $E[W_K]$ in the exact value of the mean delay in an finite buffer, while $E'[W_K]$ is it's approximation based on the results obtained for the corresponding buffer of infinite capacity, similarly P(overflow) means the exact value, while P'(overflow) in its approximation based on the results obtained for the corresponding buffer of infinite capacity.

The probability that a new customer is not accepted into a system of finite buffer capacity K, shortly , the probability of overflow, can be approximated in two ways :

$$P'_1(\text{overflow}) = P(N_\infty \geq K) \qquad (6a)$$

or

$$P'_2(\text{overflow}) = P(N_\infty > K) \qquad (6b)$$

where $P(N_\infty \geq K)$ is the probability of having K or more customers in the queueing system with infinite buffer capacity, and $P(N_\infty > K)$ is the probability that there are more than K customers in that system.

A justification for using the former approximation explained above is that, the event $(N_\infty \geq K)$ can be associated with finding the corresponding queueing system of capacity K full, while $P(N_\infty > K)$ can be associated with the proportion of time when the corresponding queueing system of finite capacity K would be overflowed.

Similarly we formulate two approximations of means delay in finite buffers, by using results obtained for queueing system of infinite buffering capacity :

$$E'_1(W_K) = E(W_\infty) * P(N_\infty \leq K) \qquad (7a)$$

or

$$E'_2(W_K) = \frac{\sum_{i=0}^{K} i\, P(N_\infty = i)}{\lambda\,[1 - P'(\text{overflow})]} \qquad (7b)$$

In eq. (7b), either approximation (6a) or (6b) can be used.

The first approximation can be interpreted as the average delay in the infinite buffer system, multiplied by the proportion of time that the infinite buffer system finds itself having K or less customers. This approximation, is obtained assuming that, in the steady state theory which says that, in the steady state, the probability of system being in a particular state is the same as the proportion of time that the system is in that particular state.

The second approximation is obtained from the Little's formula for the average delay in a finite buffer system. The mean queue length is replaced by its approximation (using probabilities of states for infinite buffer system) as well as replacing P(overflow) by its approximation.

Chapter Three

Queueing Systems with Individual Arrivals

All queueing systems in this chapter, are analysed assuming that :

- Customers arrive individually
- Service facility serves one customer after customer following FIFO order
- The arrival and departure rates of customers don't change with time and are independent on the number of customers in the system
- In finite buffer system, customers arriving when the buffer are not accepted to the system and must leave

3.1 M/M/1/∞ and M/M/1/K Queueing Systems Comparison

The first comparison is between M/M/1/∞ and M/M/1/K queueing systems. Both queueing system have simple analytical formulas, note that M/M/1/K is the only finite system on which analytical formulas can be obtained.
Those simple queueing systems can be analysed using so-called 'birth-and-death' processes. Using this method an arrival of a customer is considered as a birth , while a customer's departure is considered as a death (for derivations of the formulas, discussed in this chapter see [4]).

Given the birth rates λ and the death rates μ, we have a birth and death process as described in the diagram below, where each node represents the state of the system (the number of customers present in the system) :

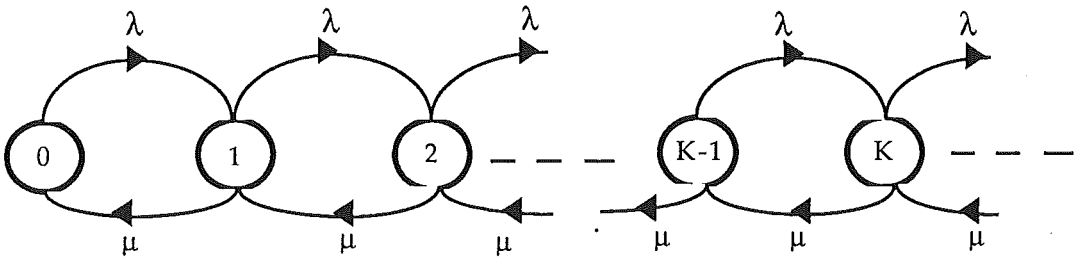


Fig. 2-1 Simple Birth-And-Death Process for M/M/1/∞ Queueing System

The probability of the infinite buffer system is in state *n* is

$$P[N_\infty = n] = (1 - \rho)\,\rho^n \qquad (8)$$

The same probability for the system of finite buffer capacity K :

$$P(N_K = n) = \begin{cases} \dfrac{(1-\rho)\,\rho^{n}}{1-\rho^{K+1}}, & \text{for } \lambda \neq \mu \\ \dfrac{1}{K+1}, & \text{for } \lambda = \mu \end{cases} \qquad (9)$$

where ρ is the utilisation factor. Thus we get the following two approximation of P(overflow)

$$P'_1(\text{overflow}) = P(N_\infty \geq K) = \rho^{K} \qquad (10a)$$

and

$$P'_2(\text{overflow}) = P(N_\infty > K) = \rho^{K+1} \qquad (10b)$$

While the average delay for buffer with infinite buffer capacity and buffer with finite capacity, respectively are :

$$E[W_\infty] = \frac{\rho}{(1-\rho)\lambda} \qquad (11)$$

and

$$E[W_K] = \frac{1}{\lambda} \frac{\rho\,(1-\rho^{K+1}) - (K+1)\,\rho^{K+1}(1-\rho)}{(1-\rho)(1-\rho^{K+1} - (1-\rho)\,\rho^{K+1})} \qquad (12)$$

Thus $E[W_K]$ can be approximated by $E'[W_K] = E[W_\infty] * P[N_\infty \leq K]$, which is equal to

$$\frac{1}{\lambda} \frac{\rho}{1-\rho} (1-\rho^{K+1}) \qquad (13)$$

Comparison Results

In Fig. 3.1 to Fig. 3.3 we have some graphs showing the comparison between the exact values of the system waiting time in M/M/1/K queueing system and their approximated values obtained from the formula (13), here we use the average value of service rate = 1.

It is obvious that as the system's buffer capacity grows, the approximation is getting more and more accurate. This effect is due to the fact that as the buffer limit increases the system buffer appears to be 'limitless', thus its behaviour is getting more and more like those of the infinite buffer systems. For any system's buffer capacity, the curve for the infinite case always greater or equal to the values for finite buffer case, and the difference is getting smaller for higher values of utilisation factor. Thus, $E'[W_K]$ always overestimates the mean delay of finite buffer system with some "safety margin".

But when the (infinite) queueing system are heavily loaded, then this approximation of the average system delay is very inaccurate. For example in fig. 3.2 and K=5, when $\rho = 0.5$, then this over-estimation is more than 50% the exact value. But again, as the system's buffer capacity increases this approximation improves, and for K = 100, the approximated values are practically equal to the exact values.

We expect that as the system utilisation gets larger, if the probability of overflow increases. This happens because, when the rates of arrival get higher, and on average, there would be more customers present in the system, thus the probability of overflow will increase as well.

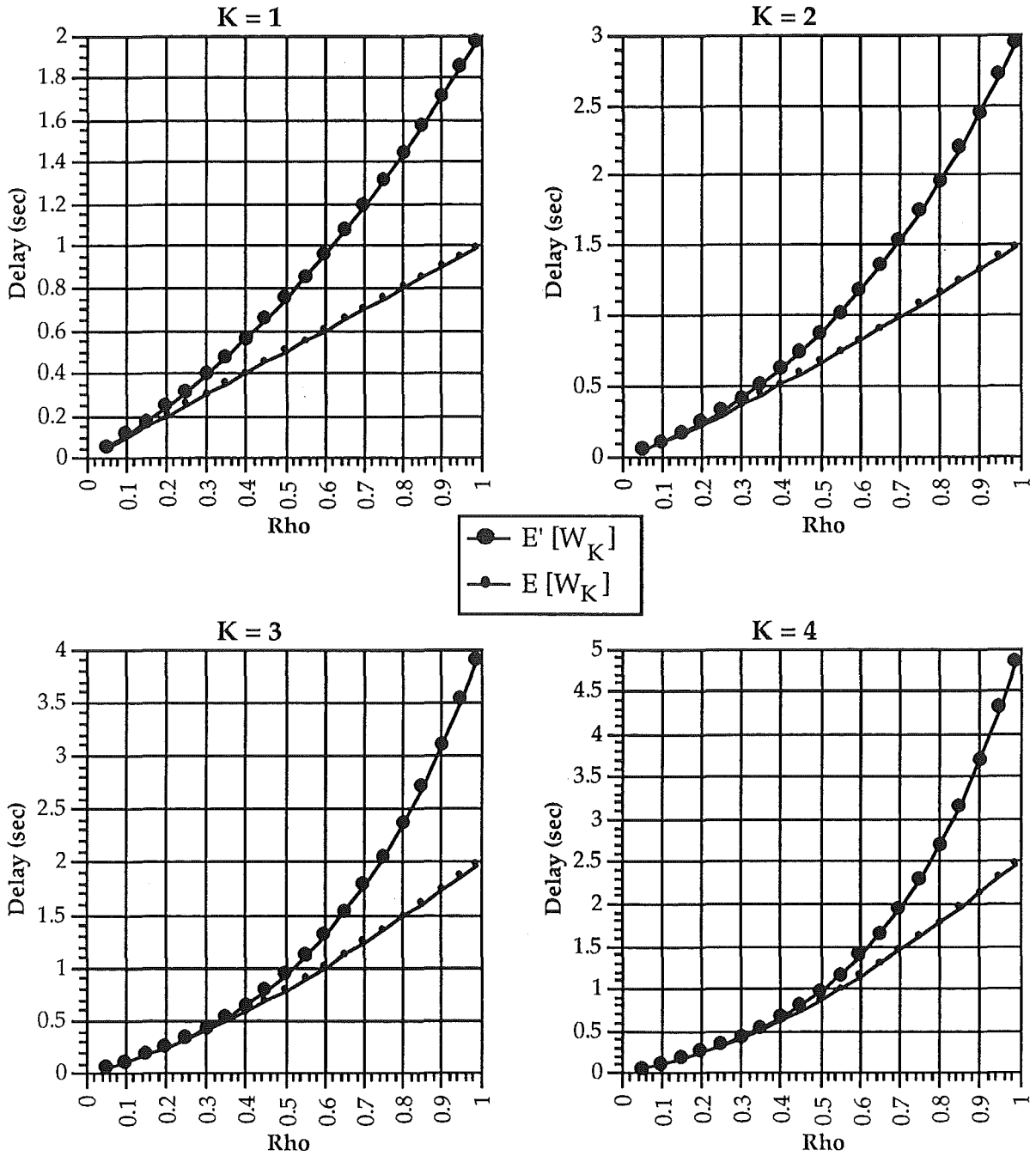
The probability of overflow approximated by $P(N_\infty \geq K)$ is always bigger than both the exact value P(overflow) and the $P(N_\infty > K)$ approximation. In case of $P(N_\infty > K)$ there is a region of ρ where this probability is smaller than of P(overflow). Of course, it's highly undesired property of any approximation of P(overflow), as it could lead to a wrong selection of system's buffer capacity. However at higher level of utilisation, $P(N_\infty > K)$ gives values greater than P(overflow) and smaller than $P(N_\infty \geq K)$, thus for higher values of ρ, $P(N_\infty > K)$ is a better approximation than $P(N_\infty \geq K)$, because it gives more accurate values, and yet, still with a safety margin. The limit value of ρ in the region where $P(N_\infty \geq K)$ approximates better P(overflow) than $P(N_\infty \geq K)$ is at about 0.5. The table 3.1 gives that critical values of ρ separating the region where $P(N_\infty \geq K)$ approximates P(overflow) better - from the region where $P(N_\infty > K)$ approximates better.

Table 3.1 The Values of ρ for that $P(N_\infty \geq K) = P(N_\infty > K)$
(for M/M/1 Queueing System)

System Size (K)	Critical Values of ρ (±0.001)	System Size (K)	Critical Values of ρ (±0.001)
1	0.619	7	0.501
2	0.544	8	0.501
3	0.519	9	0.501
4	0.509	10	0.501
5	0.505	20	0.501
6	0.503	100	0.500

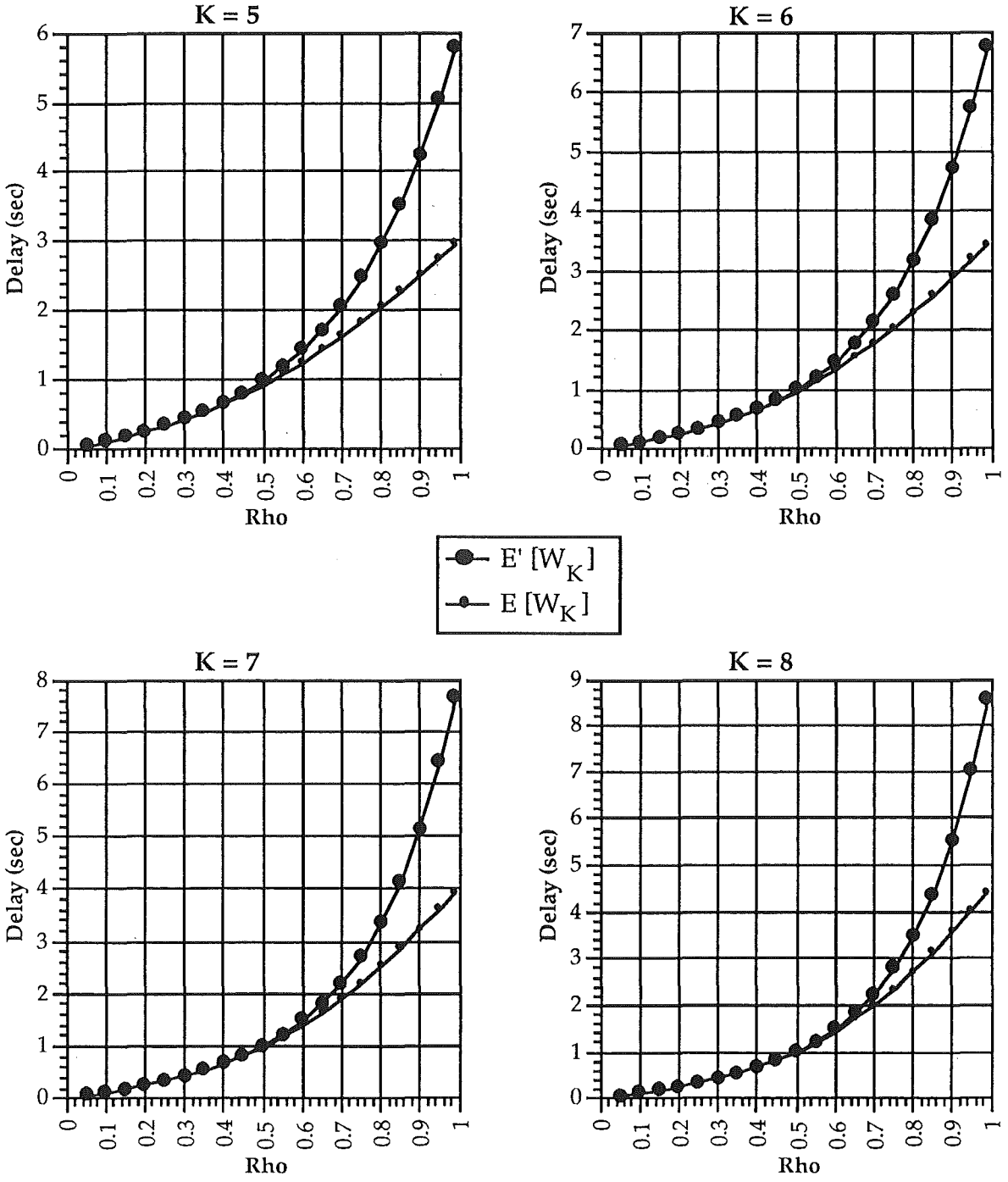
Note that for system's buffer capacity of 1, the crossing point happens when utilisation factor equal to 0.6, this somewhat deviates from the general trend of the crossing point for M/M/1 queueing system, this behaviour is considered 'normal', because here we have a queueing system with infinite queue buffer capacity being used to approximate queueing system with no space for queueing at all, so the result of this approximation is irrelevant to our interest.

Mean Waiting Time in M/ M/ 1/ K System
Comparison between
$E(W_K)$ and $E'(W_K)$



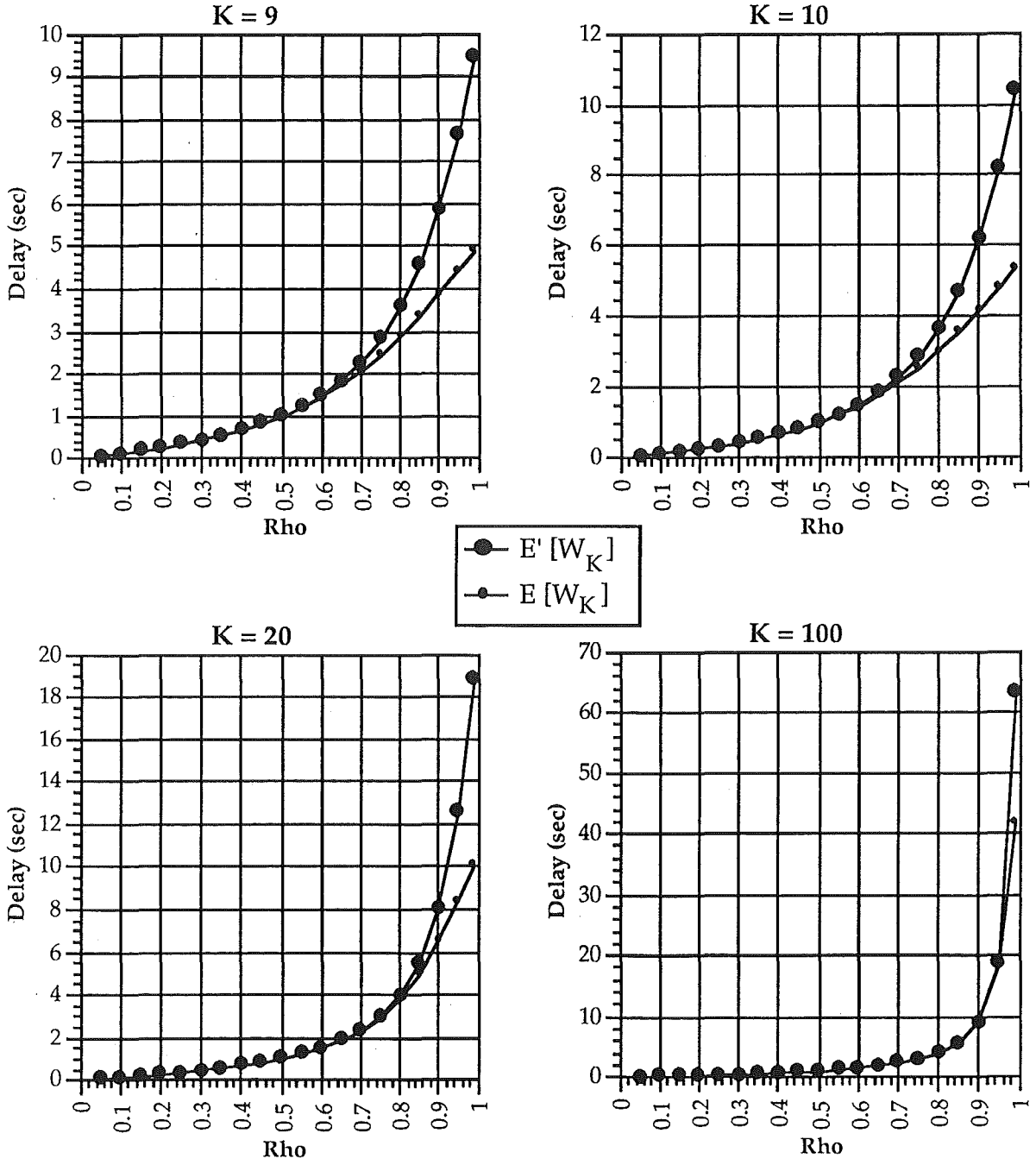
Graph 3.1

Mean Waiting Time in M/ M/ 1/ K System
Comparison between
$E(W_K)$ and $E'(W_K)$



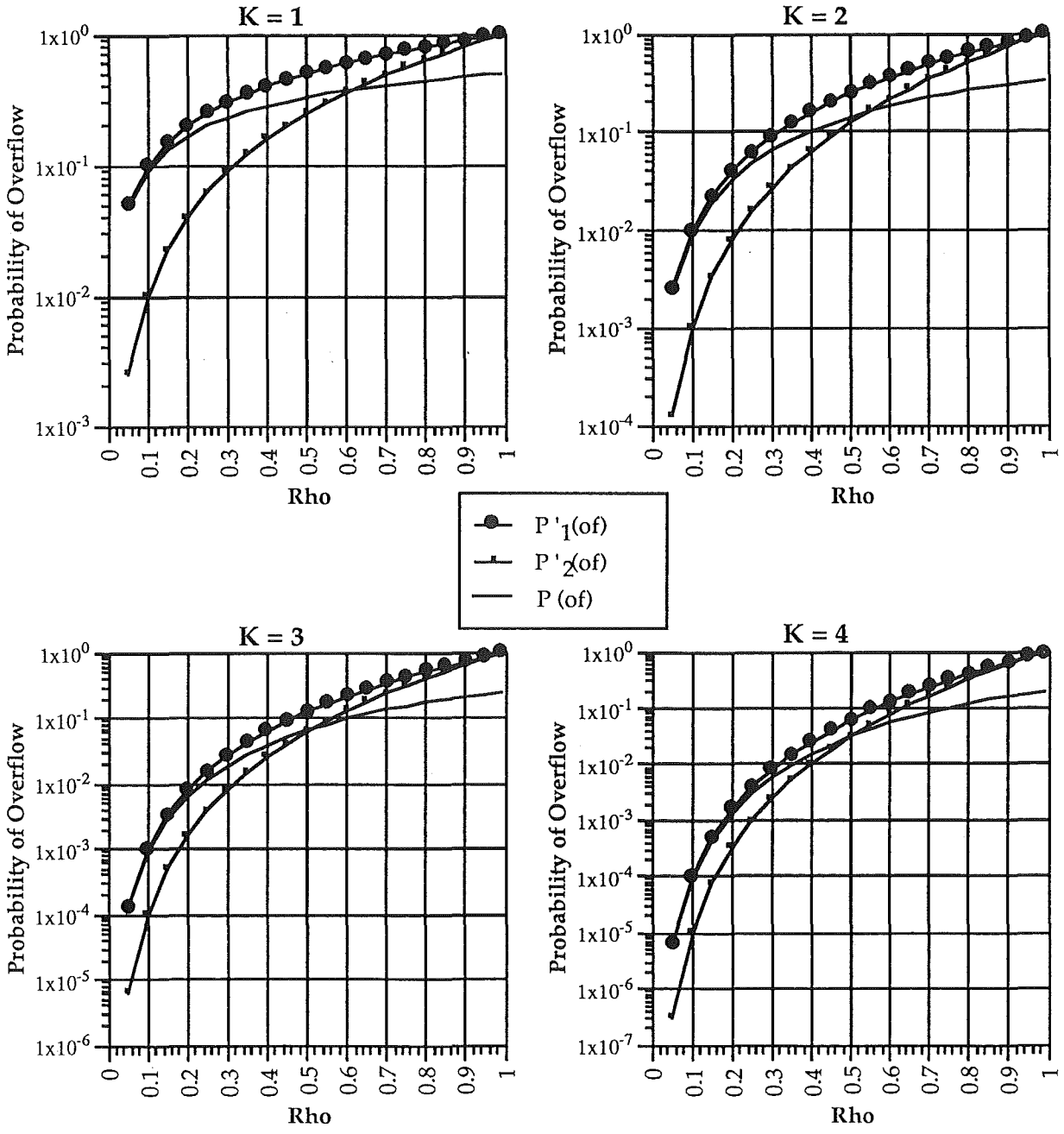
Graph 3.2

Mean Waiting Time in M/ M/ 1/ K System
Comparison between
$E(W_K)$ and $E'(W_K)$



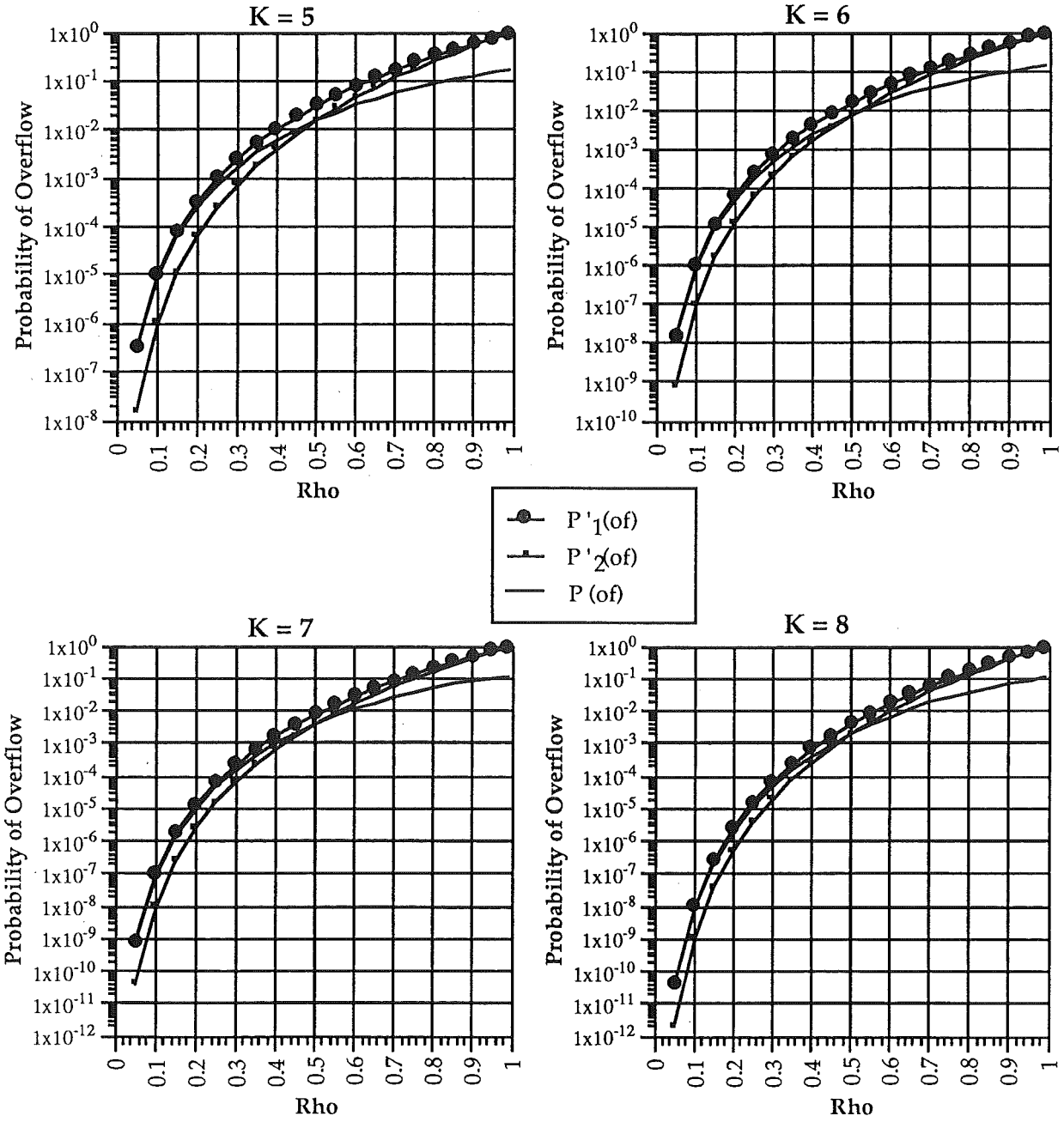
Graph 3.3

Probability of Overflow in M/ M/ 1 / K
Comparison between
P(overflow), P_1' (overflow), P_2' (overflow)



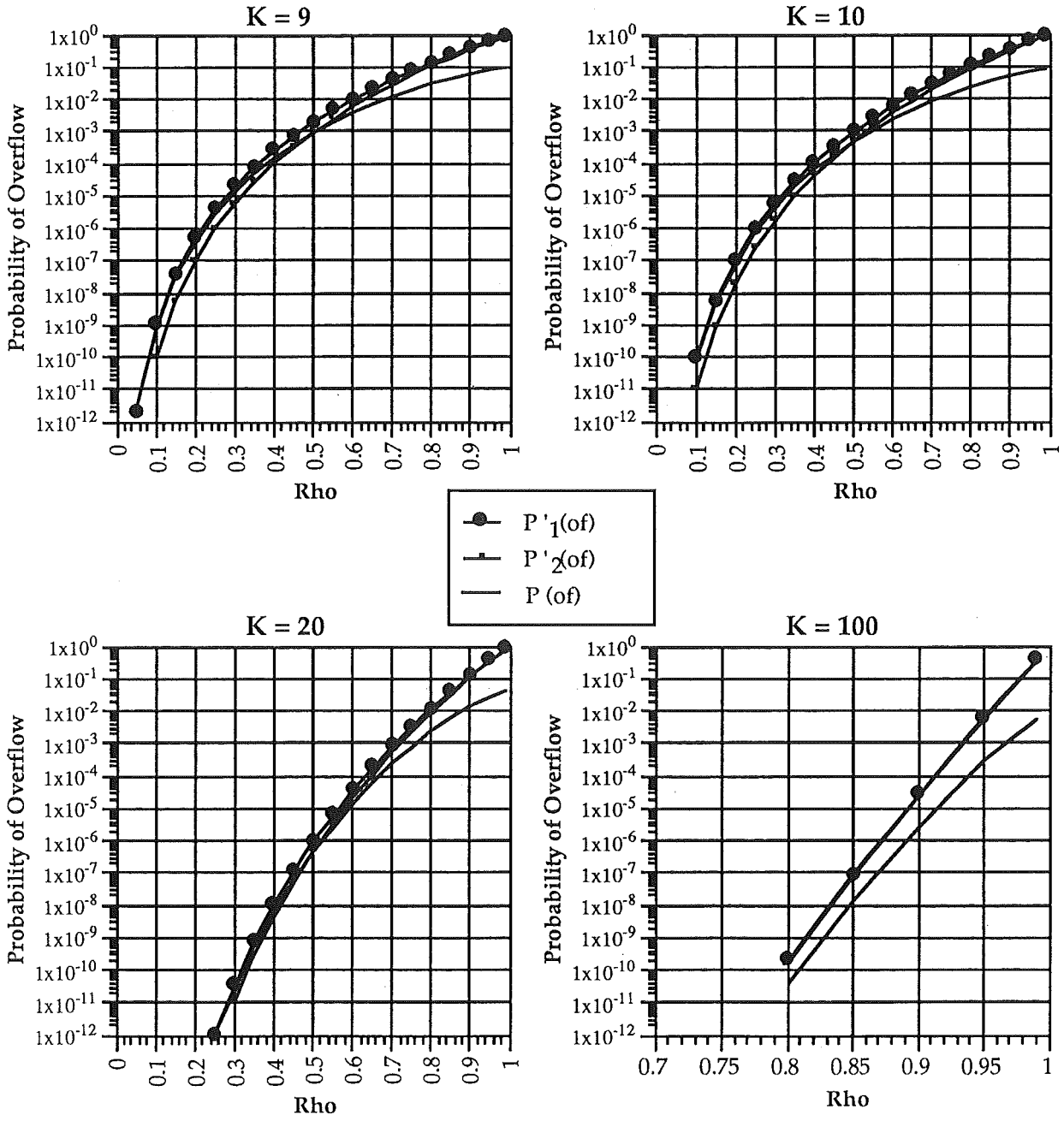
Graph 3.4

Probability of Overflow in M/ M/ 1 / K
Comparison between
P(overflow), P_1' (overflow), P_2' (overflow)



Graph 3.5

Probability of Overflow in M/ M/ 1 / K
Comparison between
P(overflow), P_1' (overflow), P_2' (overflow)



Graph 3.6

3.2 M/D/1/∞ and M/D/1/K Queueing Systems Comparison

These queueing systems need more elaborate analysis. The results could be obtained by means of techniques developed for M/G/1/∞ and M/G/1/K queueing system.
The most commonly used method of analysis is known as the 'Embedded Markov Chain' method. Following that we view the system just after every departure (service completion), unlike in case of the birth-and-death model, where we can observe the system at any instant time.
To find the steady state probabilities of an embedded Markov chain, we must solve a set of equations that can be expressed in a matrix form. For example, assuming the system capacity of K :

$$\pi = \pi P \quad \text{and} \quad \sum_{i=0}^{K} \pi_i = 1 \tag{14}$$

where π_i is the steady state probability P(N = i), $\pi = (\pi_1, \pi_2, \pi_3, ..., \pi_K)$, and P is the transition matrix. A transition matrix is a (KxK) matrix whose elements, are the probabilities that n customers arrive during one service period, denoted by a_n's , thus

$$a_n = P[\text{Arrival} = n] = \int_0^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \, dW_s[t], \quad n = 0, 1, 2.. \tag{15}$$

where $W_s[t]$ is the service time distribution, and λ is the arrival rates.

Without going further into details of the formulas' derivations, we just say that the probability of the system being in the state n is π_n, thus P[N = n] = π_n , This fact is proven by Gross and Harris [2]. The formulas used to approximate the probability of overflow are as follows :

$$P'_1(\text{overflow}) = P(N_\infty \geq K) = 1 - \sum_{i=0}^{K-1} P(N_\infty = i) \tag{16a}$$

and

$$P'_2(\text{overflow}) = P(N_\infty > K) = 1 - \sum_{i=0}^{K} P(N_\infty = i) \tag{16b}$$

For the average system delay :

$$E'[W_K] = \frac{\sum_{i=0}^{K} i\, P(N_\infty = i)}{\lambda(1 - P(N_\infty \geq K))} \qquad (17)$$

and the exact formula for average system delay is

$$E[W_K] = \frac{\sum_{i=0}^{K} i\, P(N_K = i)}{\lambda\,[1 - P(N_K = K)]} \qquad (18)$$

Comparison Results

The quality of the approximations, can be seen in Fig 3.7 to Fig 3.12, and here we use $W_s = 1$.

When K (system capacity) = 1, and assuming that the average service rate is 1, the average system waiting times are all equal to one, for all values of utilisation factor (see Fig 3.7, for K=1). The reason is that, when K=1, we don't actually have any queue in front of the server, so if a customer arrives when the service facility is busy, the customer is rejected right away, and leaves the system. On the other hand, if the service facility is idle, any new customer is served straight away with a constant service rate equal to 1 (consistent with the assumption above), thus all serviced customers have the average system delay equal to the service rates which is 1.

Figures 3.7 - 3.9 show that the approximation $E'[W_K]$ given by Eq.(17) always results in higher values than those produced by the exact formula (18), and we see that as the system's buffer capacity gets larger, again the two curves merge together closer and closer, therefore, the bigger the system's buffer capacity is, the more accurate the approximation is.

For example, for K = 100 (see Fig. 3.9, K=100), all the values of $E'[W_K]$ are virtually equal to the exact values. Even when K = 7, for utilisation factor or $\rho = 0.9$, the approximation $E'[W_K]$ is only off by 5%, which can be considered accurate enough, taking into account a usual safety margin.

Comparing the results we obtained for M/M/1 and M/D/1 queueing systems, we see that for both queueing systems $P(N_\infty \geq K)$ always produces greater values $P(N_\infty > K)$ and the exact values of P(overflow).

Similarly, there is a utilisation region where $P(N_\infty > K)$ approximation is better than $P(N_\infty \geq K)$, by the same reasons as in the case of M/M/1 queueing system. The table below shows us the critical values of ρ for applying $P(N_\infty \geq K)$ and $P(N_\infty > K)$. For K=100, the crossing point is not displayed on the graph, because the precision used in the computation of the probabilities was limited to 10^{-12}.

Table 3.2 The Crossing Points Between $P(N_\infty \geq K)$ and $P(N_\infty > K)$ Curves for M/D/1 Queueing System

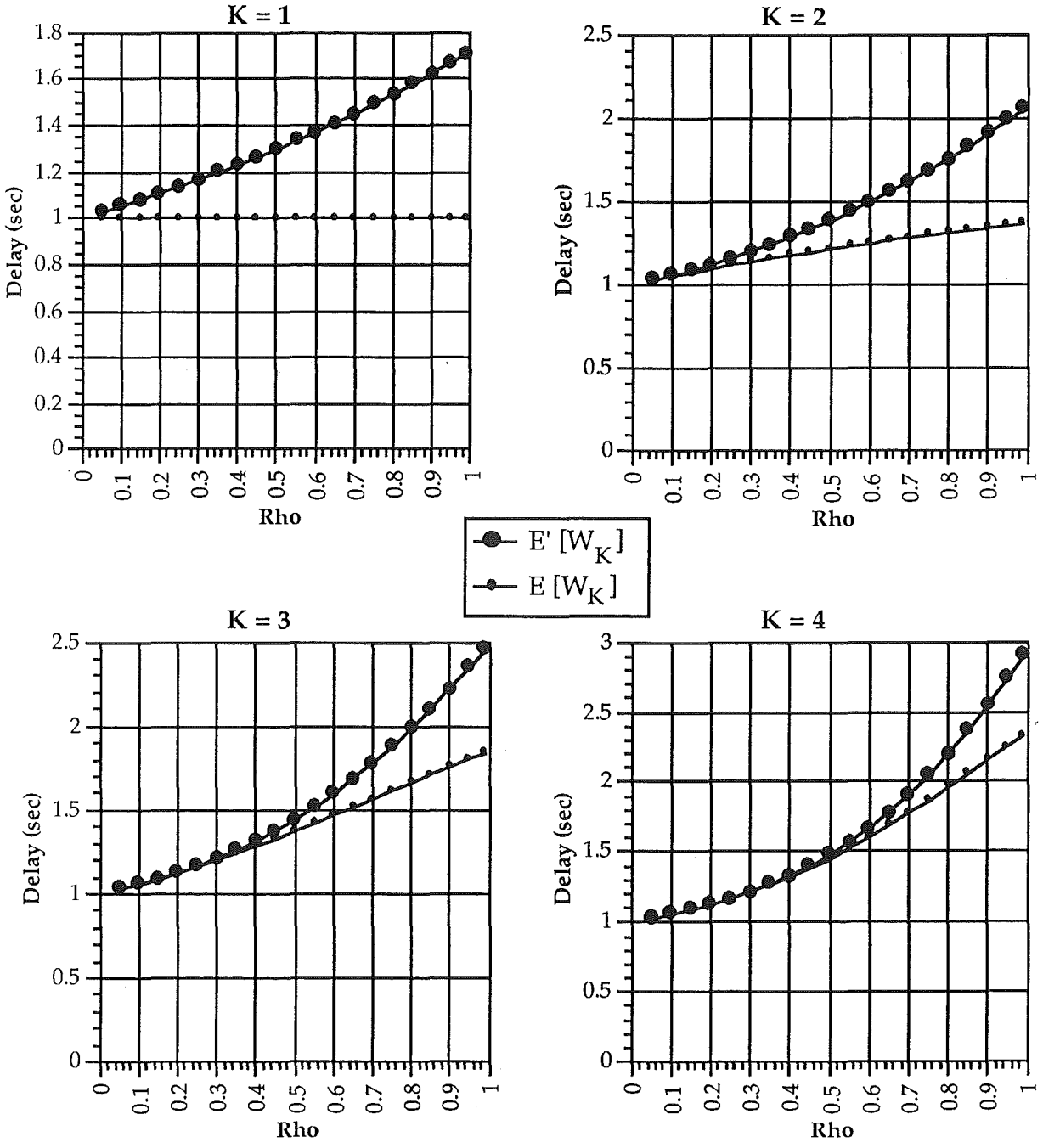
System Size K	Critical Values of ρ (±0.001)	System Size K	Critical Values of ρ (±0.001)
1	0.715	7	0.607
2	0.643	8	0.606
3	0.620	9	0.606
4	0.612	10	0.606
5	0.608	20	0.606
6	0.607	100	0.601

Conclusion

For single arrival queueing systems, M/M/1 and M/D/1, the approximations of the average system delay are reasonably good, especially for queueing system with large buffer capacity.

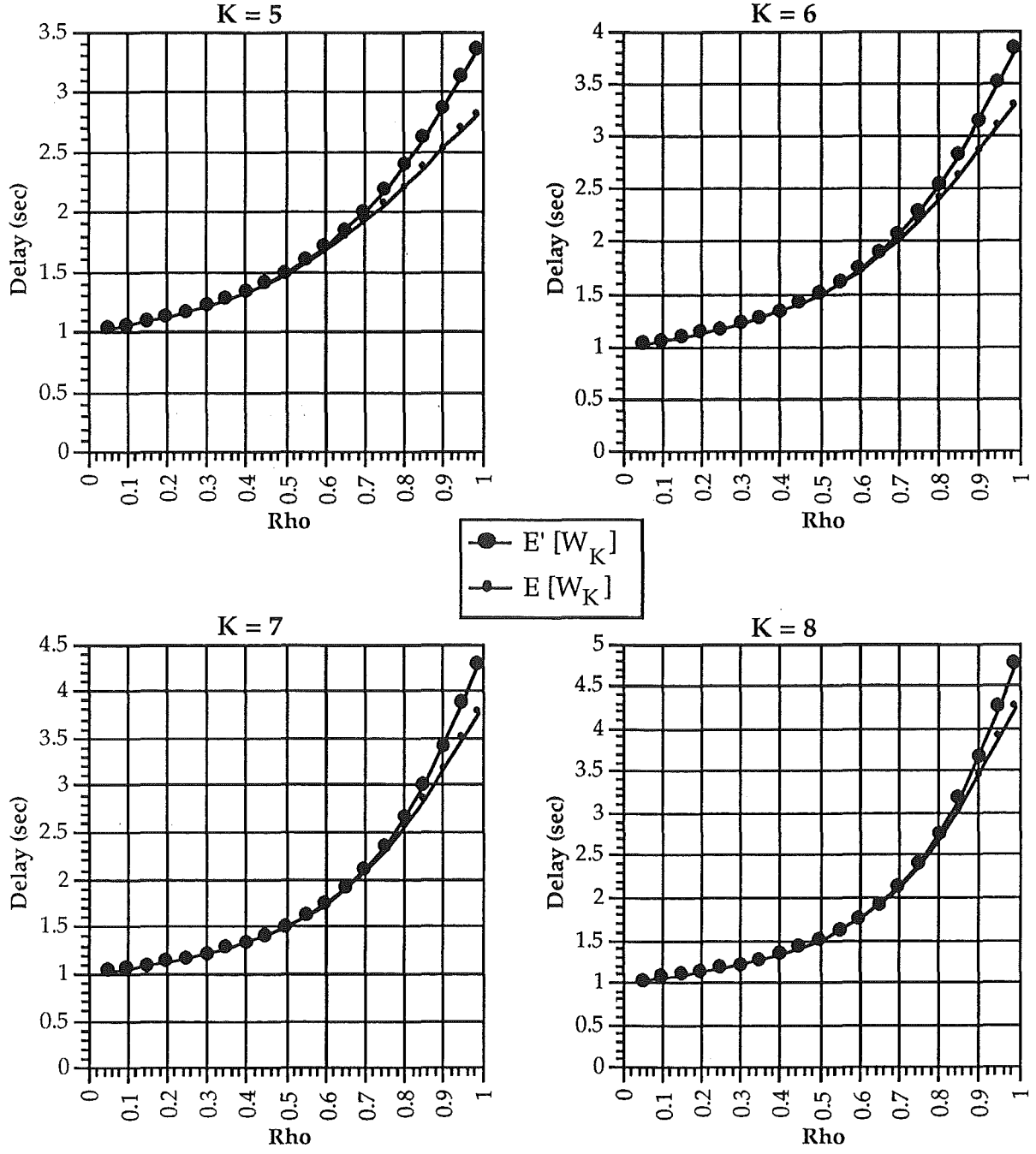
Approximations of the probability of overflow , $P(N_\infty \geq K)$ is a better approximation of P(overflow), than $P(N_\infty > K)$, for lower utilisation factors, while for higher utilisation factors the $P(N_\infty > K)$ approximates better.

Mean Waiting Time in M/ D/ 1/K System
Comparison between
$E(W_K)$ and $E'(W_K)$



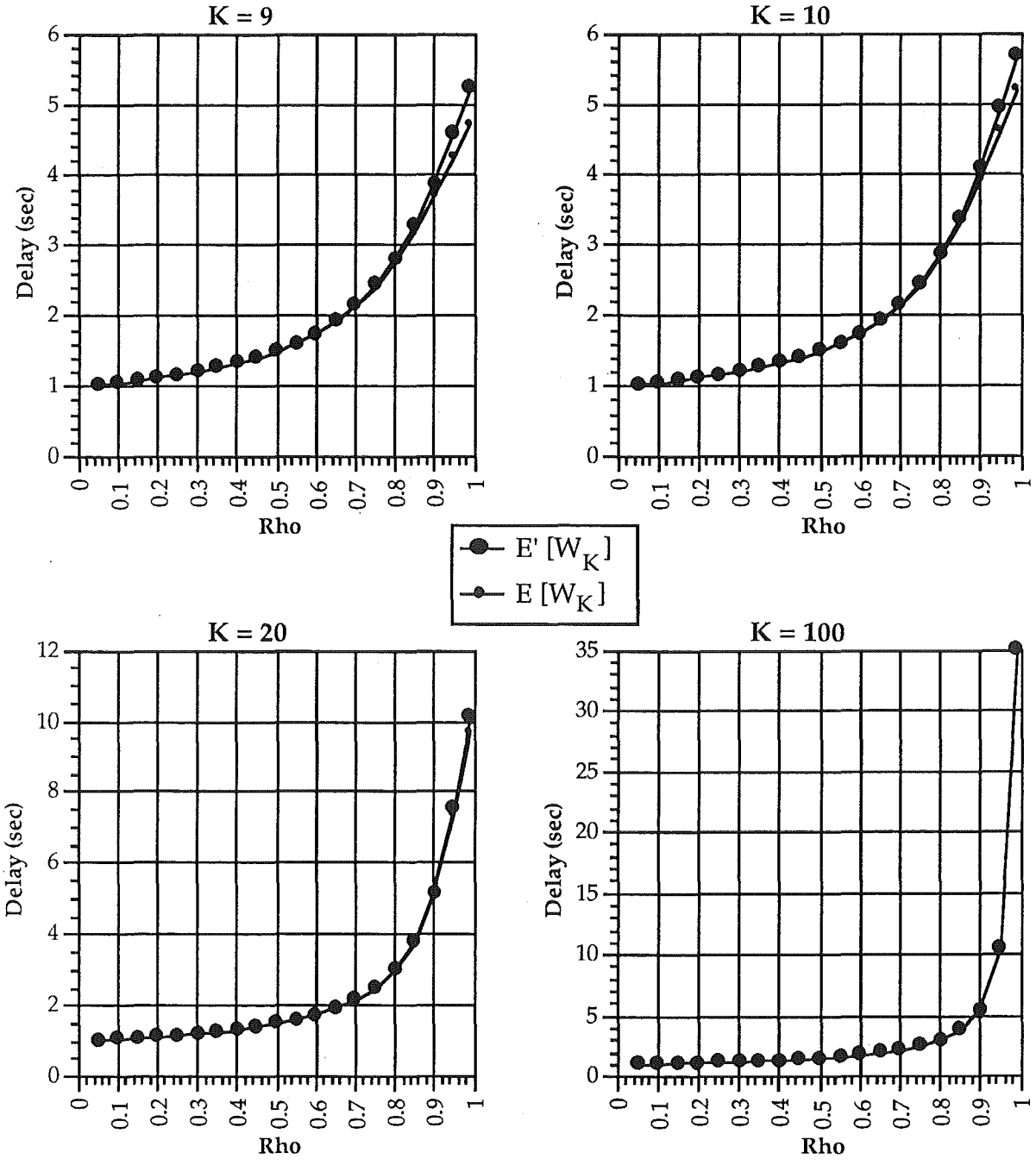
Graph 3.7

Mean Waiting Time in M/ D/ 1/K System
Comparison between
$E(W_K)$ and $E'(W_K)$



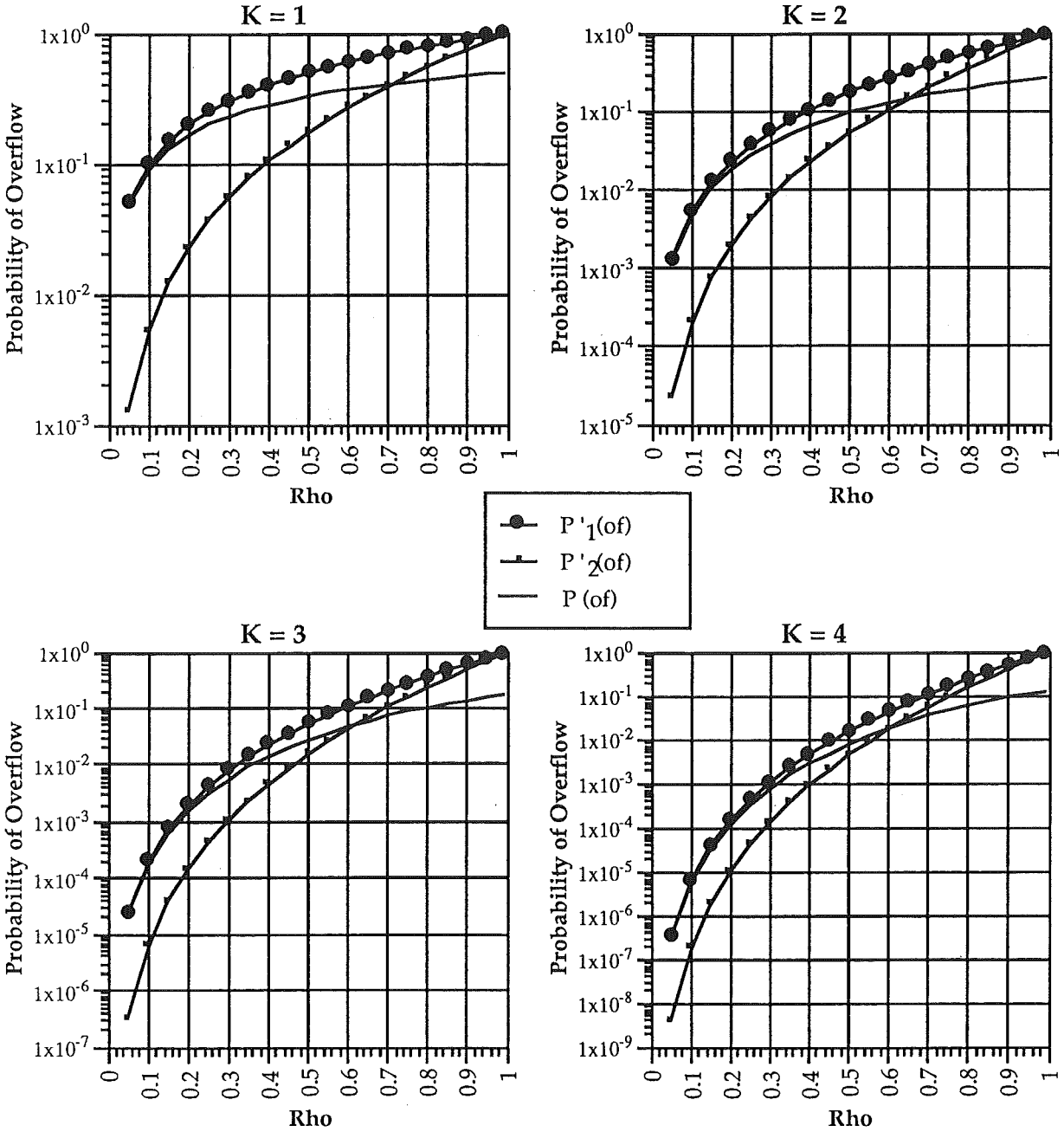
Graph 3.8

Mean Waiting Time in M/ D/ 1/K System
Comparison between
$E(W_K)$ and $E'(W_K)$



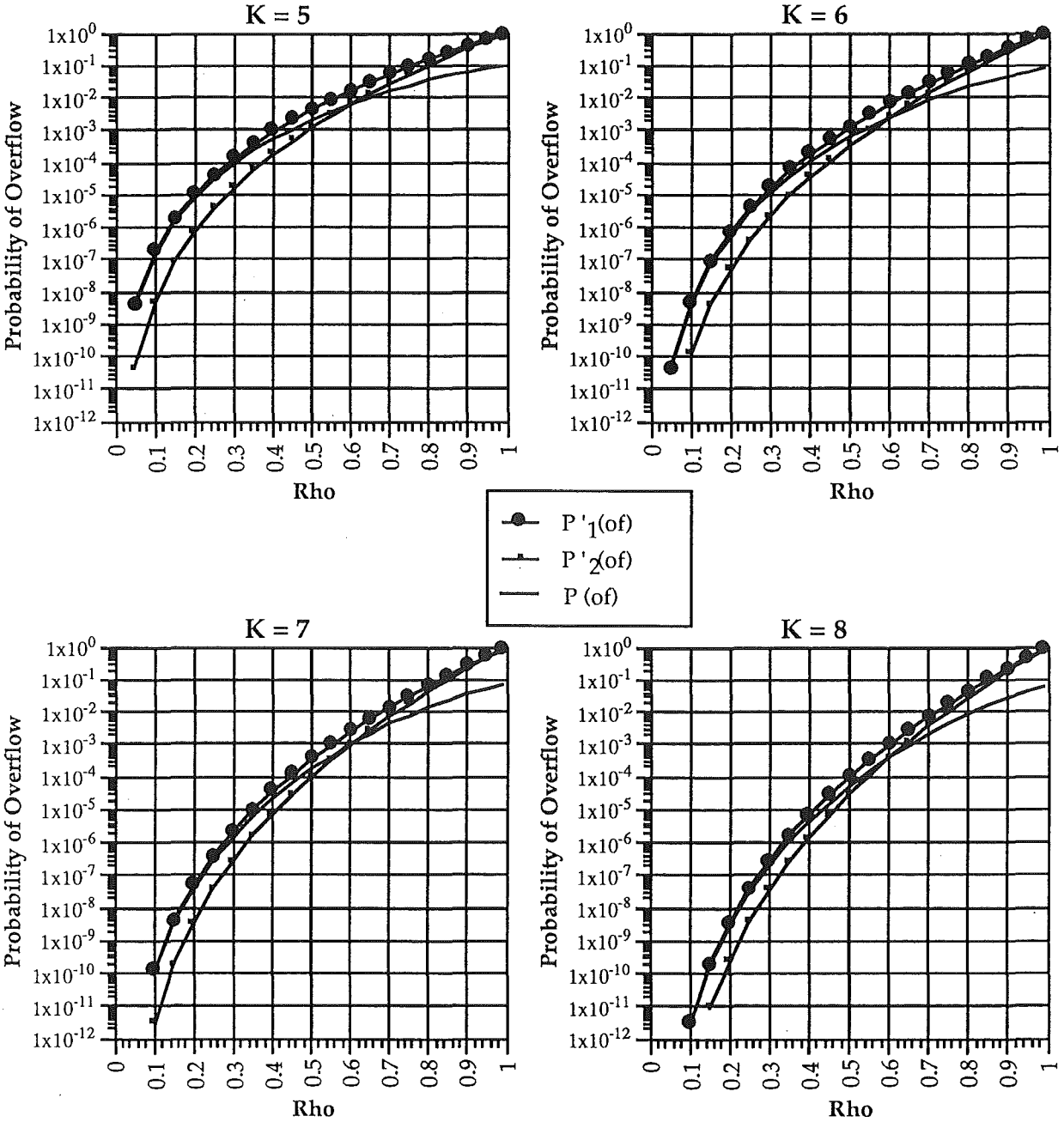
Graph 3.9

Probability of Overflow in M/ D/ 1 / K
Comparison between
P(overflow), P_1' (overflow), P_2' (overflow)



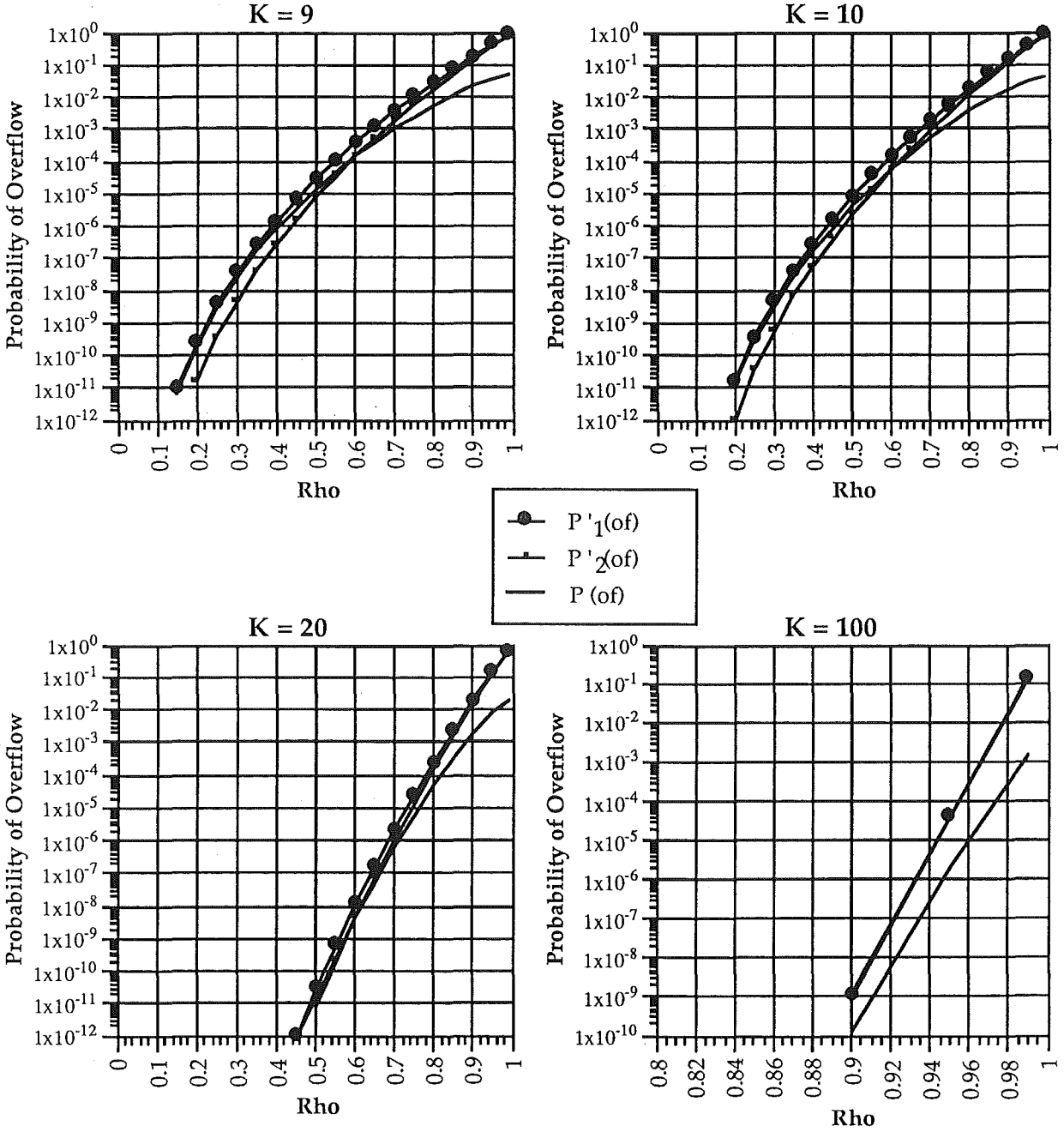
Graph 3.10

Probability of Overflow in M/ D/ 1 / K
Comparison between
P(overflow), P_1' (overflow), P_2' (overflow)



Graph 3.11

Probability of Overflow in M/ D/ 1 / K
Comparison between
P(overflow), P_1' (overflow), P_2' (overflow)



Graph 3.12

Chapter Four

Queueing Systems with Batched Arrivals

If we the assumption that customers arrive individually, then we have the $M^{(b)}/M/1$ queueing system. In this queueing system, we still have Poisson arrival stream, but the difference is that , the of customers arrive in random groups (batches), ie. more than one customer can arrive at the same instant of time.
The state diagram for this system is depicted below :

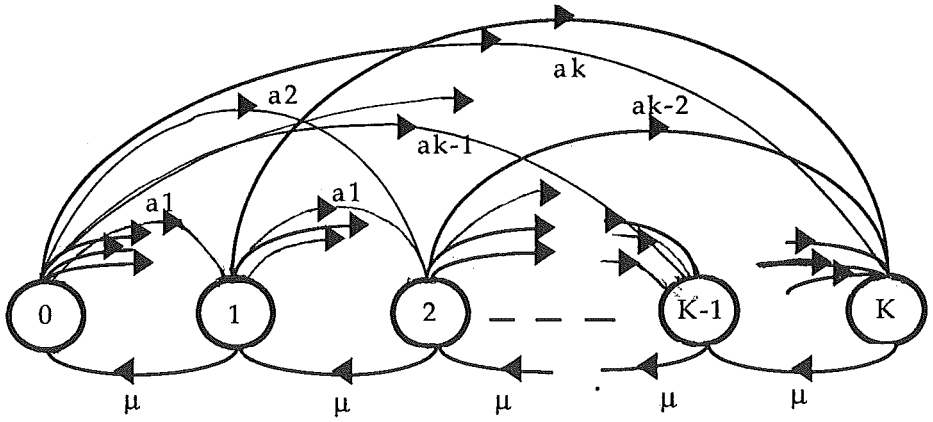


Fig 4.1 State transition diagram for $M^{(b)}/M/1/K$ queueing system

In the queueing systems with individual arrivals the transitions are only allowed to the nearest neighbours, for example from state 4 we can have a transition to state 5 or state 3, but not to, say, state 10.
On the other hand, in queueing systems with batched arrivals, multi-step transitions occur whenever batches of more than one customer arrive.
The size of an arriving batch can take any positive integer value, and we assume that the batch size *i* occurs with the probability a_i.

From the state transition diagram above, we can derive balance equations that imply the 'flow conservation' principle, see Gross and Harris [2].

Following that principle, for system with finite buffer capacity, we have balance equations as follows :

$$P[N_K = n]\ (\mu + \sum_{i=1}^{K-n} a_i) \;=\; P[N_K = n+1]\ \mu + \sum_{j=1}^{n} P[N_K = n-j\ ]\ a_j \qquad (19)$$

and since

$$P[N_K = 0] \sum_{i=1}^{K} a_i = \mu P[N_K = 1] \qquad (20)$$

thus

$$P[N_K = 1] = \frac{P[N_K = 0] \sum_{i=1}^{K} a_i}{\mu} \qquad (21)$$

where a_i is assumed to be from geometric distribution, thus $a_i = (1 - q)\, q^{i-1}$ for $(0 < q < 1)$.

Each of the probabilities P[N = n] can be computed recursively starting from N = 0, then working all the way up from N = 0 to K.

The calculation of the average system response is still using the same method as previously done for the M/D/1 queueing system (based on Little's formula). However the calculation of the probability of overflow is more complicated than as it was before for queueing systems with individual arrivals. Let us note that overflow can happen even in the state 0, if more than K new customers arrive. plus the probability of state 1 and K-1 or more customers arrive. In state 1 overflow happens if more than K-1 customers arrive, etc. The exact formula for the probability of overflow, and its approximations are listed in the Appendix A-3.

The Comparison

The comparison is done for three different value of q (parameter for the geometric distribution of the batch size), namely 0.2, 0.5, 0.8, what corresponds to the average batch sizes of 5, 2, 1.25. The results are shown in Fig. 4.1 to 4.18.

It is noticeable, that the curve for $E'[W_\infty]$ in every case where system's buffer capacity is 1, is missing. This effect is caused by the fact that P'(overflow) = 1 for all values of utilisation factor ρ and q, and we try to approximate a queueing system with no queue buffer capacity, with a queueing system with unlimited buffer capacity.

Looking at the graph for mean system's waiting time, we don't see any significant difference, for various q values except that the waiting time is higher when q is bigger (see graph 4.1, 4.2, 4.3), as an example, see the table 4.1 below, where system's buffer capacity is 6 :

Table 4.1 The effect of average batch size to the average system waiting time (in secs.)

Utilisation Fact.	Delay, q= 0.2	Delay, q= 0.5	Delay, q = 0.8
0.1	1.95	3.55	12.2
0.4	2.87	7.20	13.9
0.6	4.02	10.10	85.7

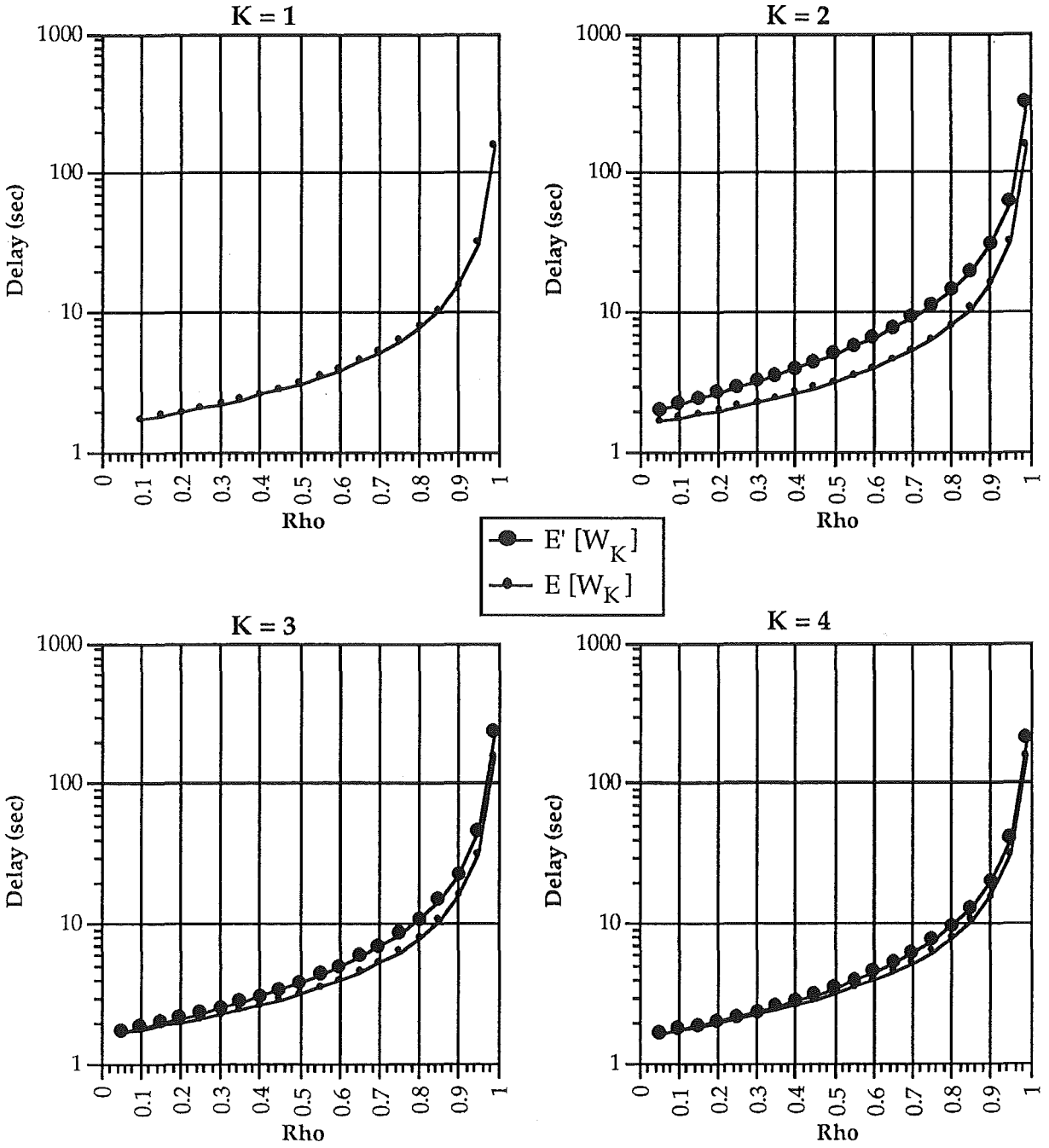
If we look at the values of P'_1(overflow) and P'_2(overflow), we see that again $P(N_\infty \geq K)$ is always bigger than both the $P(N_\infty > K)$ and the exact value, but this time the $P(N_\infty > K)$ curve never crosses the $P(N_\infty \geq K)$ curve. In fact for $q = 0.8$, the $P(N_\infty \geq K)$ is under P(overflow) curve. That fact makes us able to conclude that for the batched arrival case of $M^{(b)}/M/1$ queueing system, $P(N_\infty \geq K)$ is the better approximation of P(overflow).

Conclusion.

The average batch size of arrivals effects the performance of queueing system, in such a way, that the smaller the average batch size, the shorter the average system waiting time.

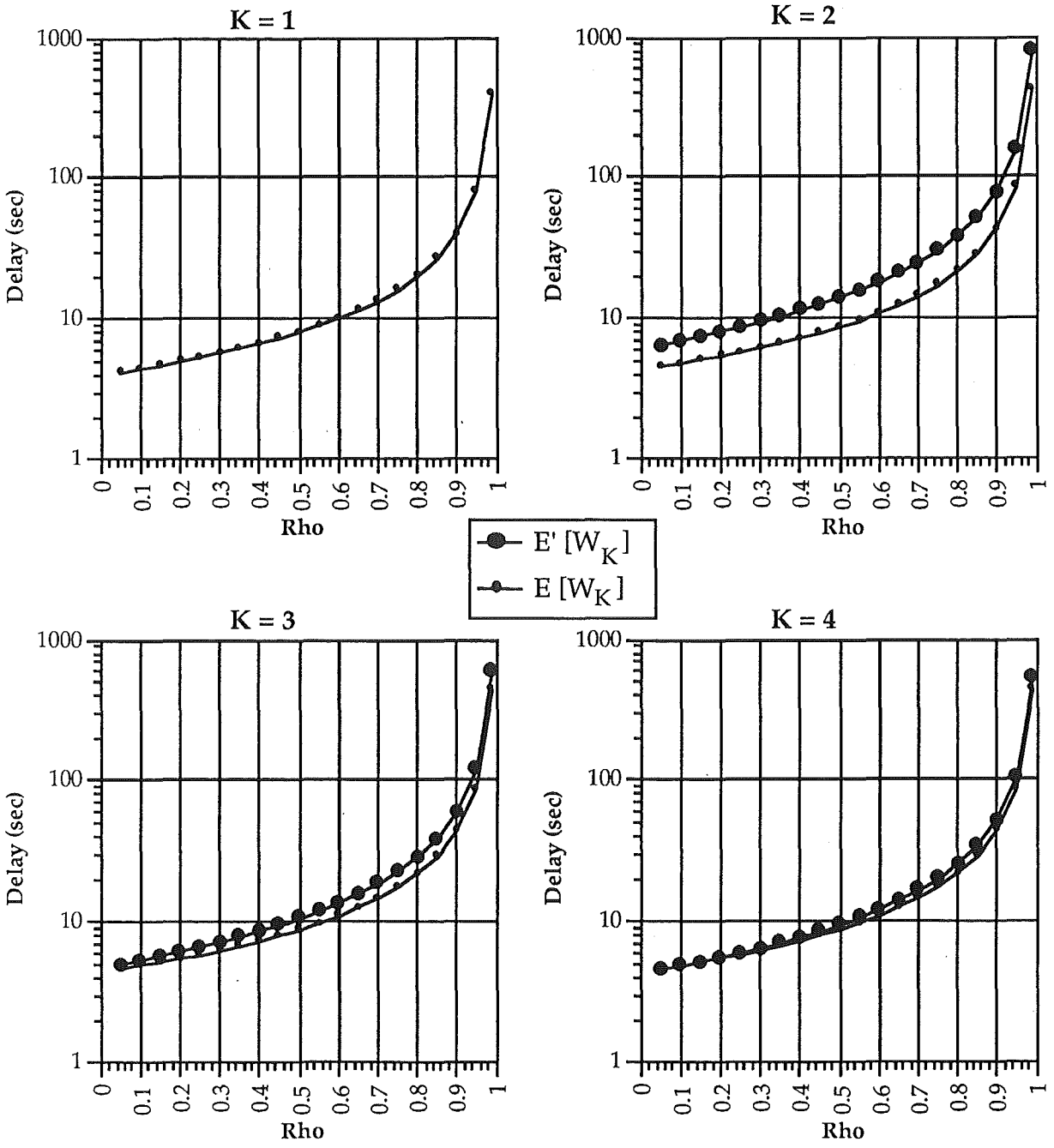
Another observation is that $P(N_\infty \geq K)$ is better approximation of the exact value P(overflow) for any range of utilisation coefficient. It's unlike the queueing system with individual arrivals, where in some range of utilisation factor, $P(N_\infty > K)$ approximates better than $P(N_\infty \geq K)$.

Mean Waiting Time in $M^{(b)}/M/1/K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.2$



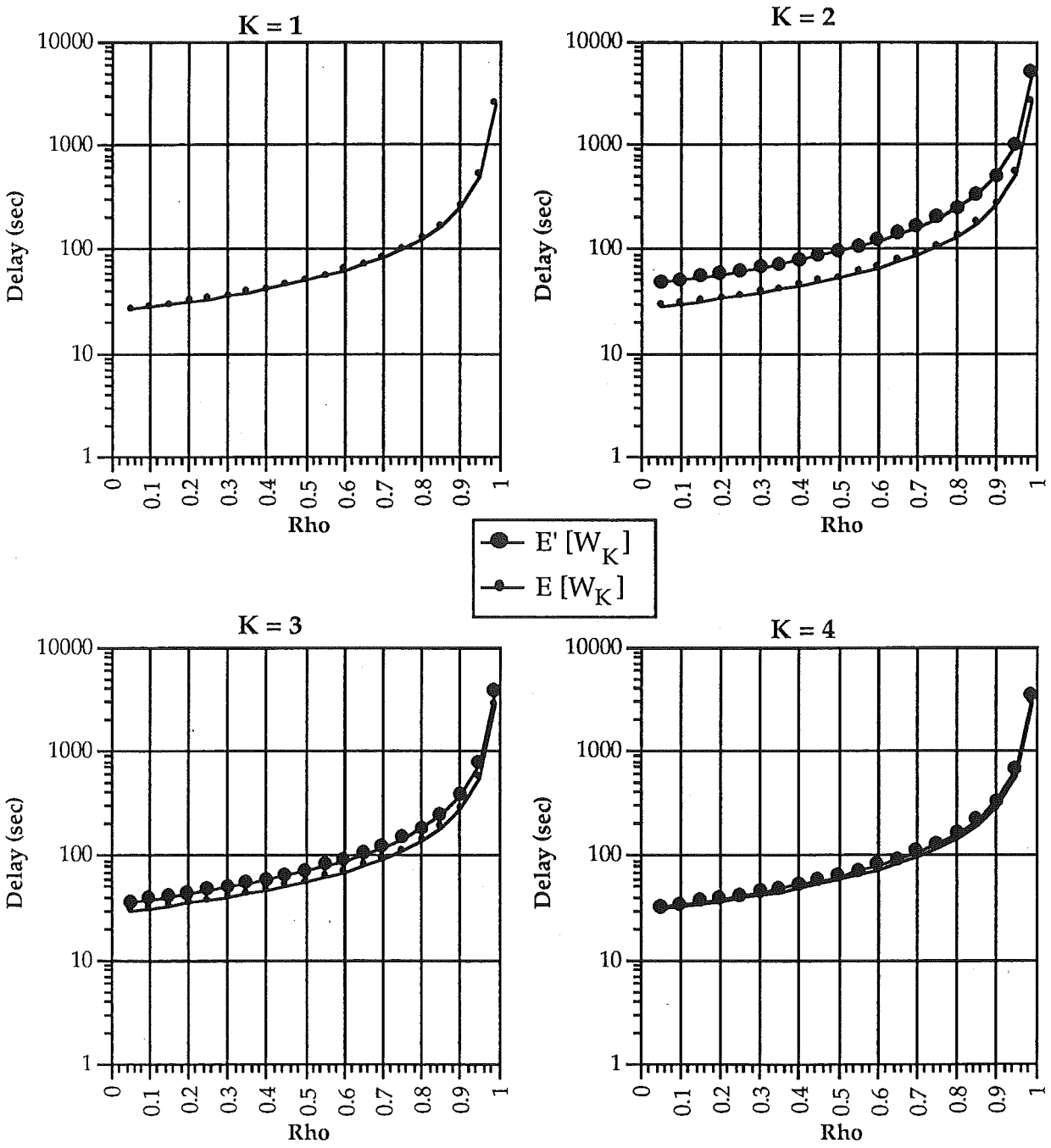
Graph 4.1

Mean Waiting Time in $M^{(b)}/M/1/K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.5$



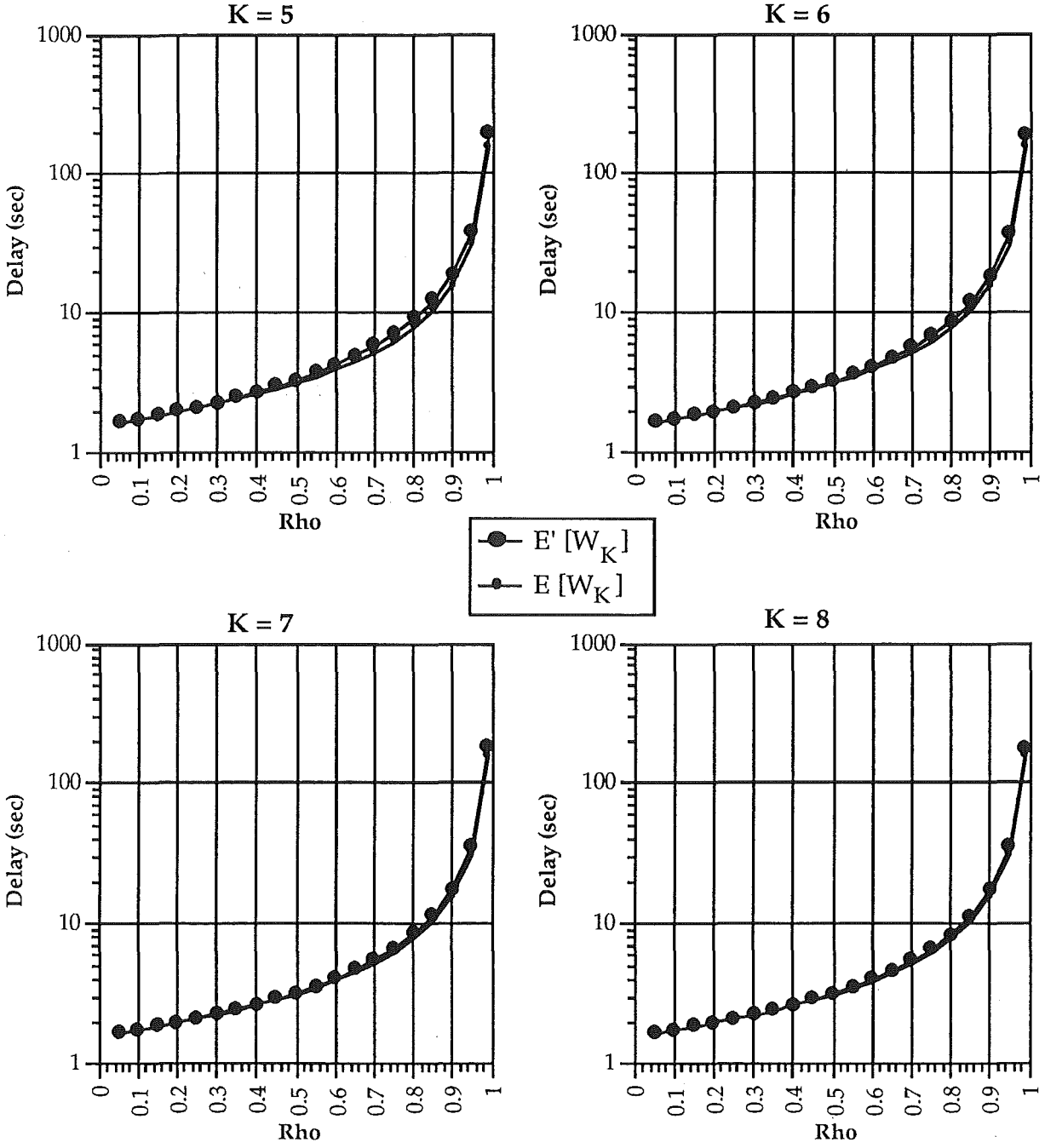
Graph 4.2

Mean Waiting Time in $M^{(b)} / M / 1 / K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.8$



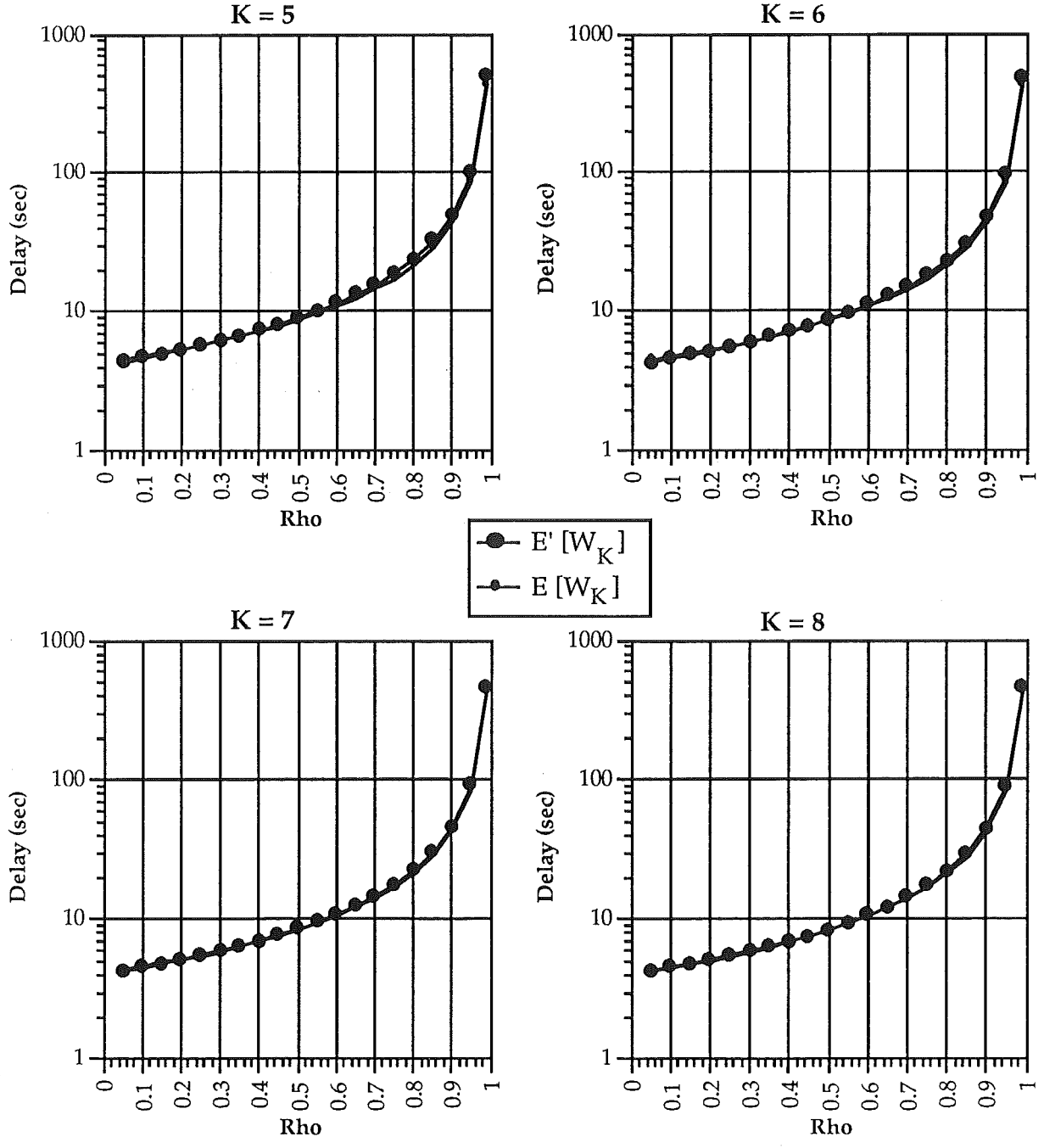
Graph 4.3

Mean Waiting Time in $M^{(b)}/M/1/K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.2$



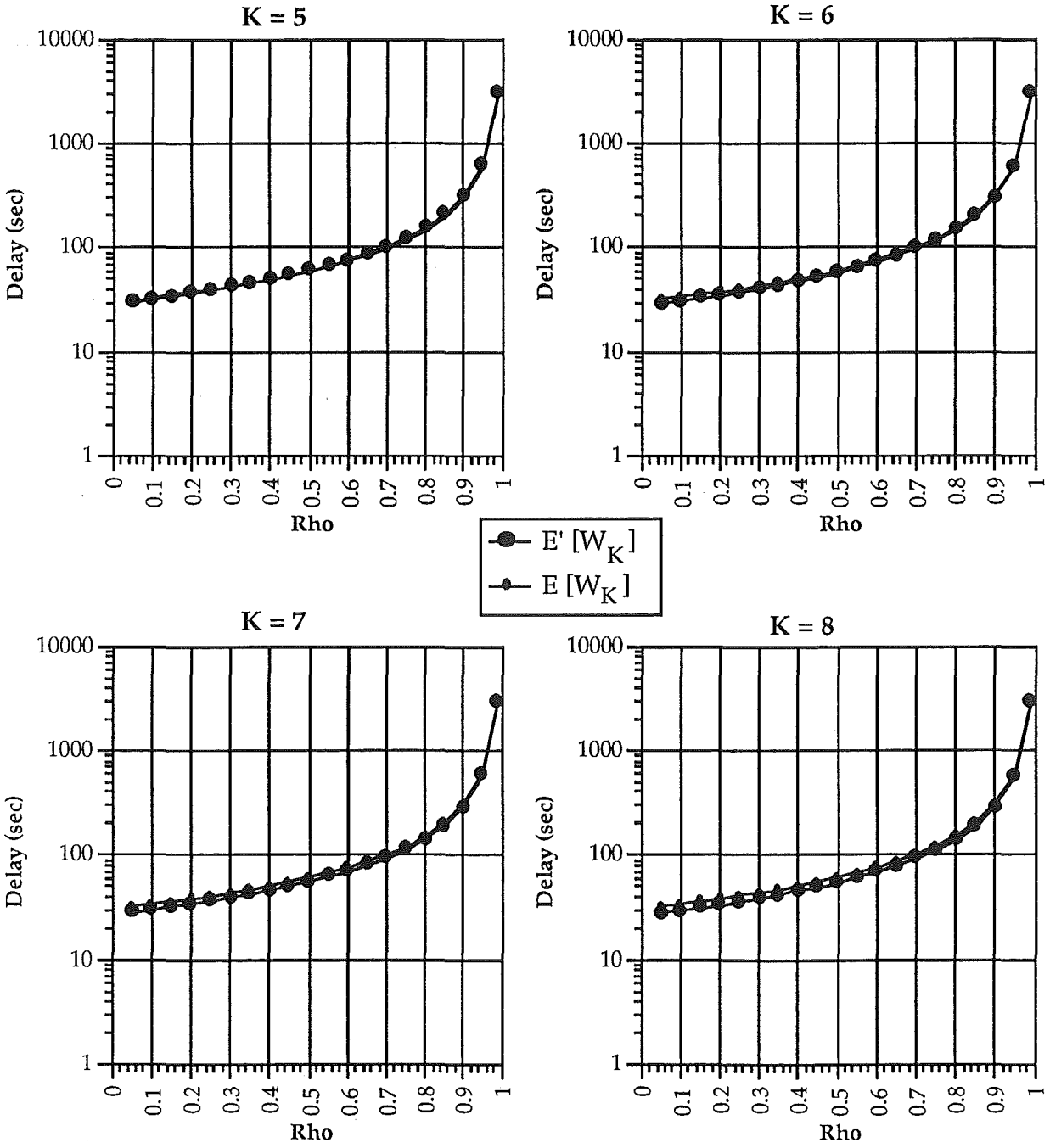
Graph 4.4

Mean Waiting Time in $M^{(b)}/M/1/K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.5$



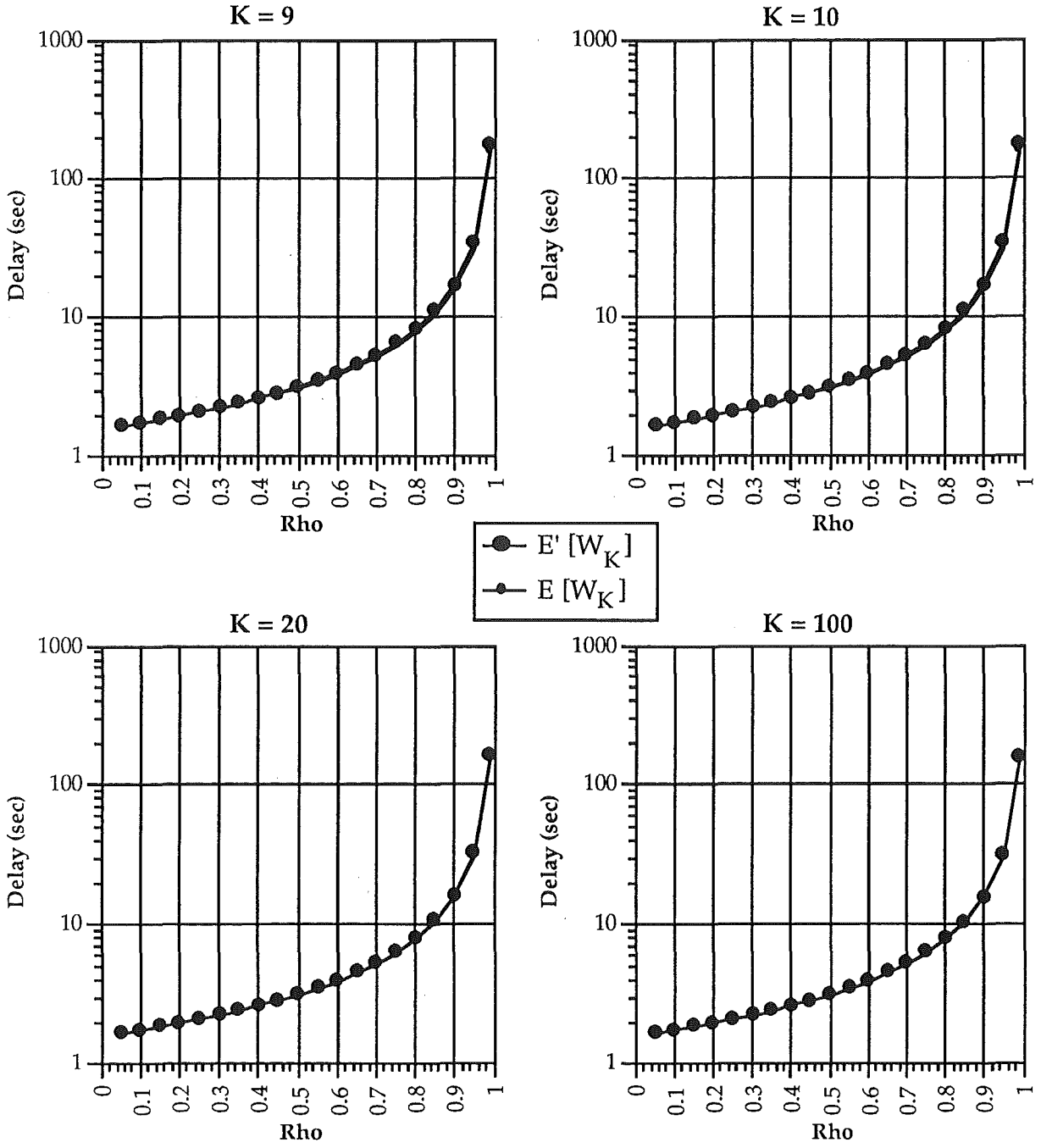
Graph 4.5

Mean Waiting Time in $M^{(b)}/M/1/K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.8$



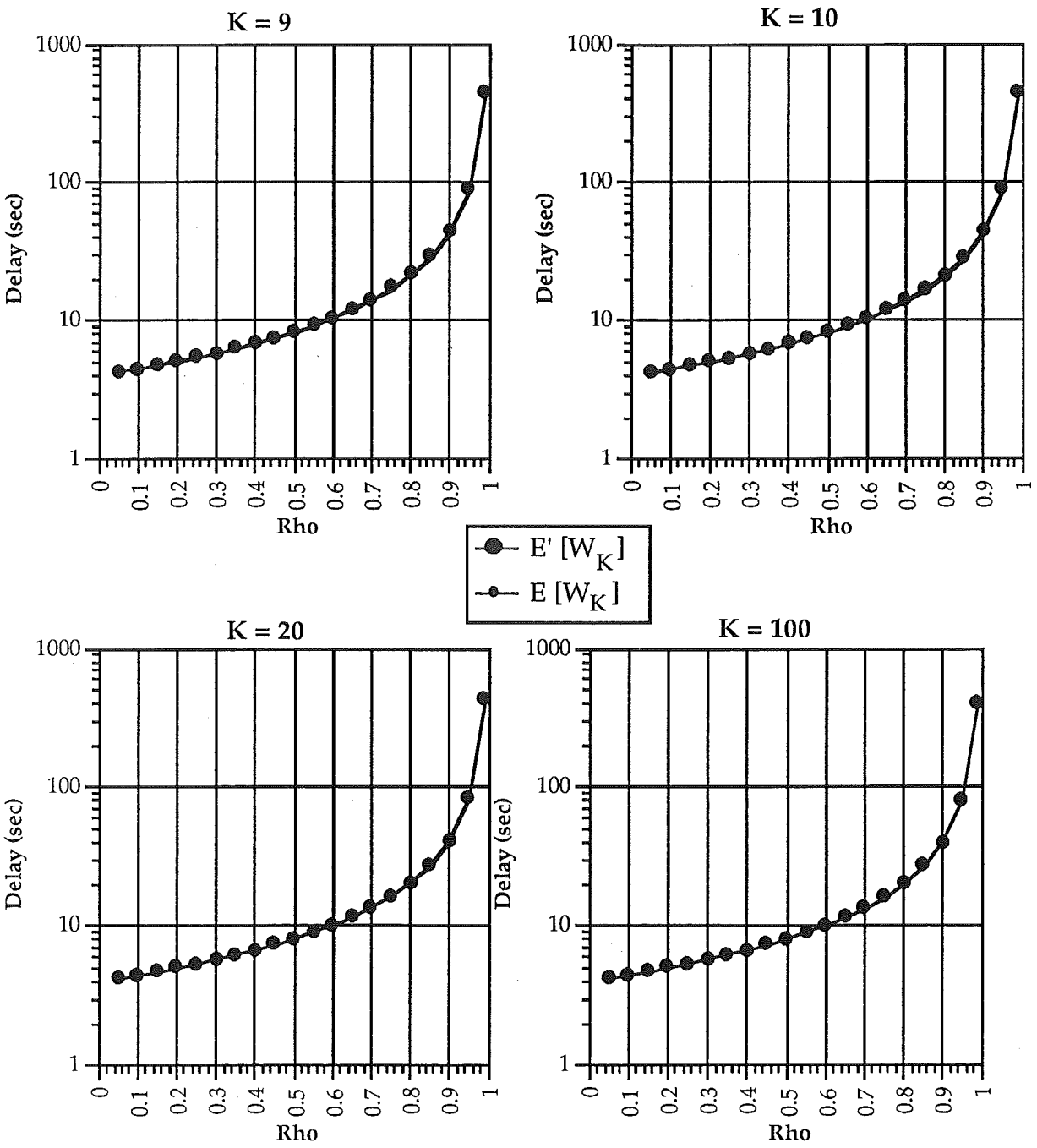
Graph 4.6

Mean Waiting Time in $M^{(b)}/M/1/K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.2$



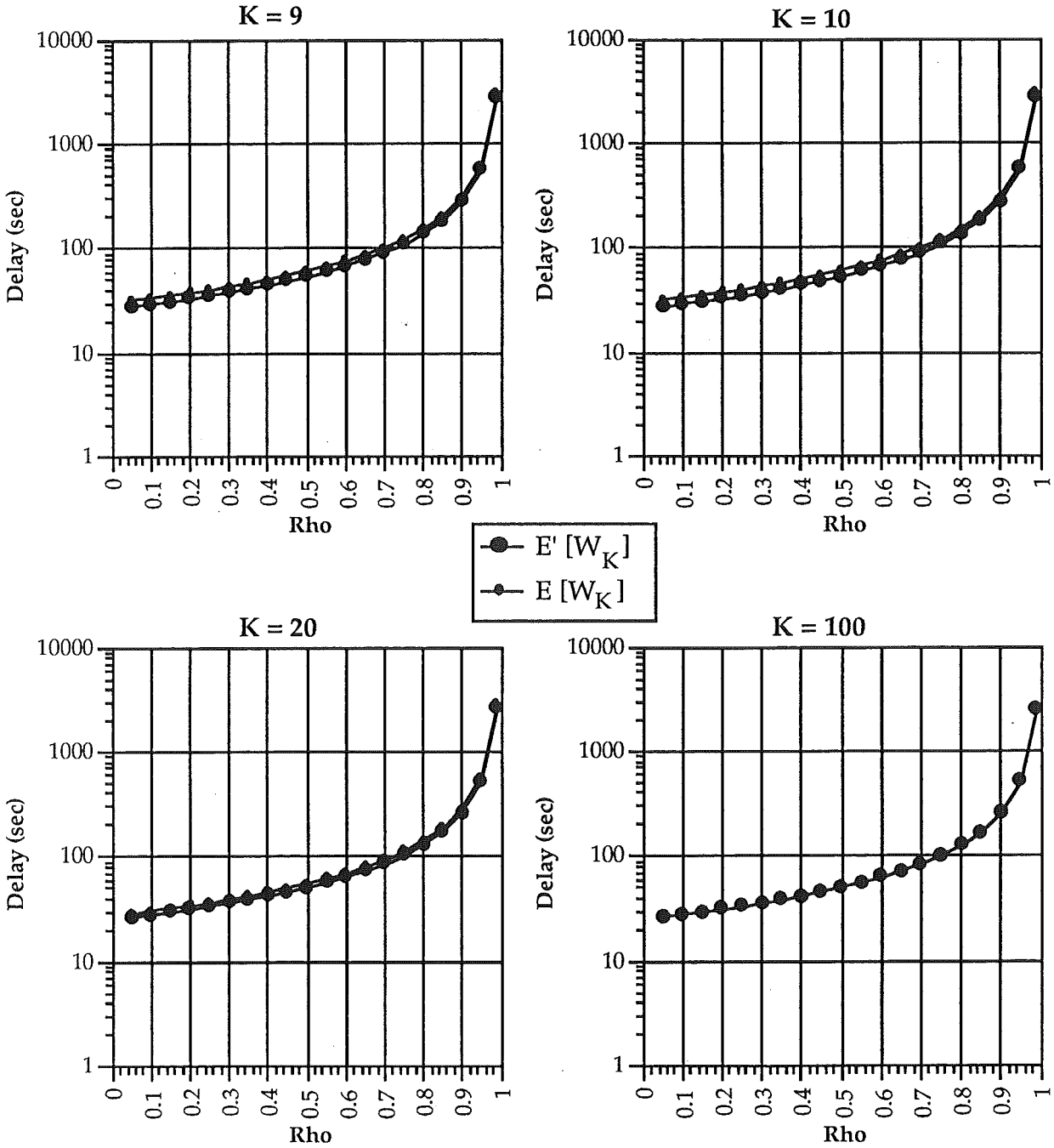
Graph 4.7

Mean Waiting Time in $M^{(b)}/M/1/K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.5$



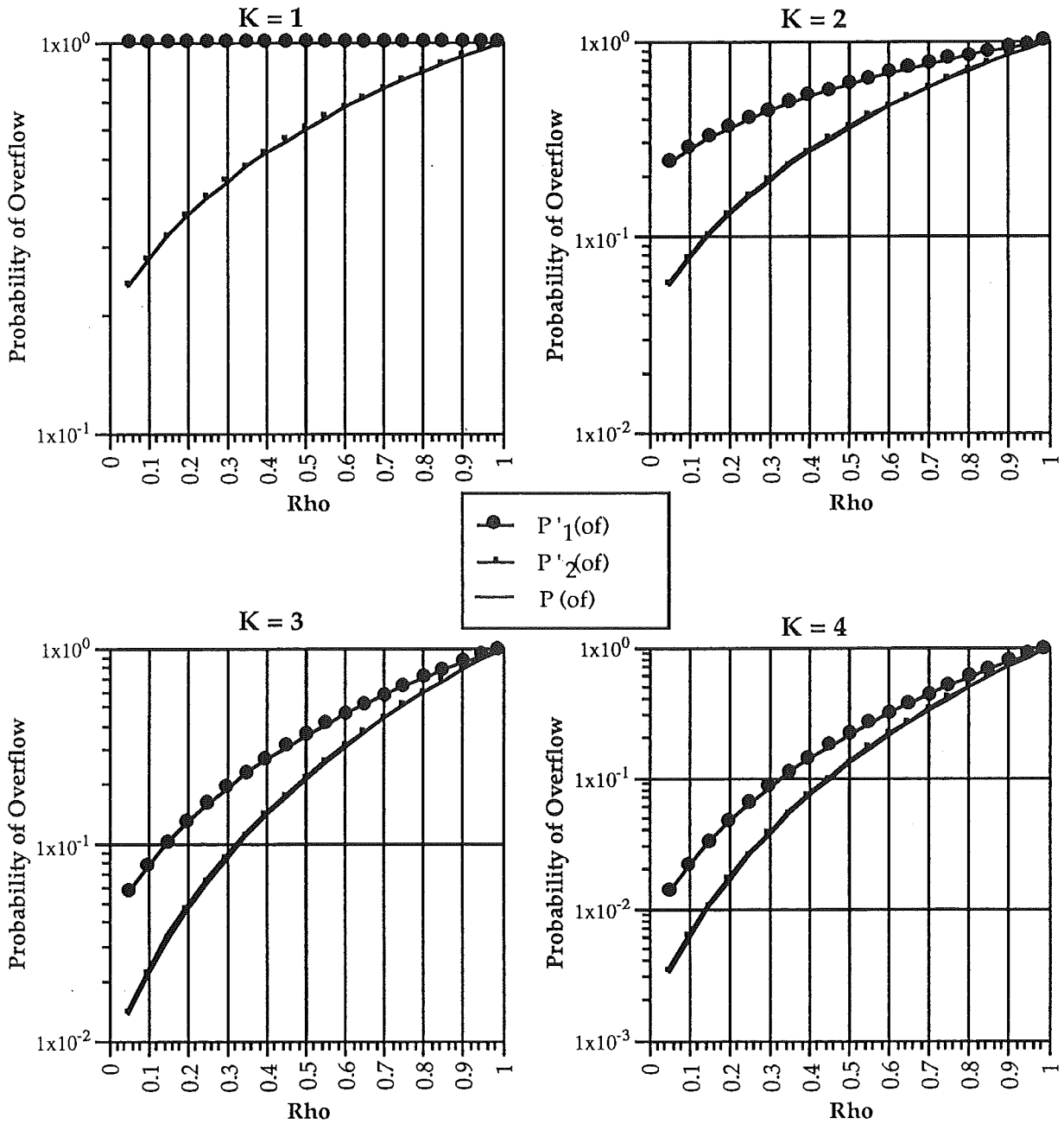
Graph 4.8

Mean Waiting Time in $M^{(b)} / M / 1 / K$
Comparison Between
$E(W_K)$ and $E'(W_K)$
$q = 0.8$



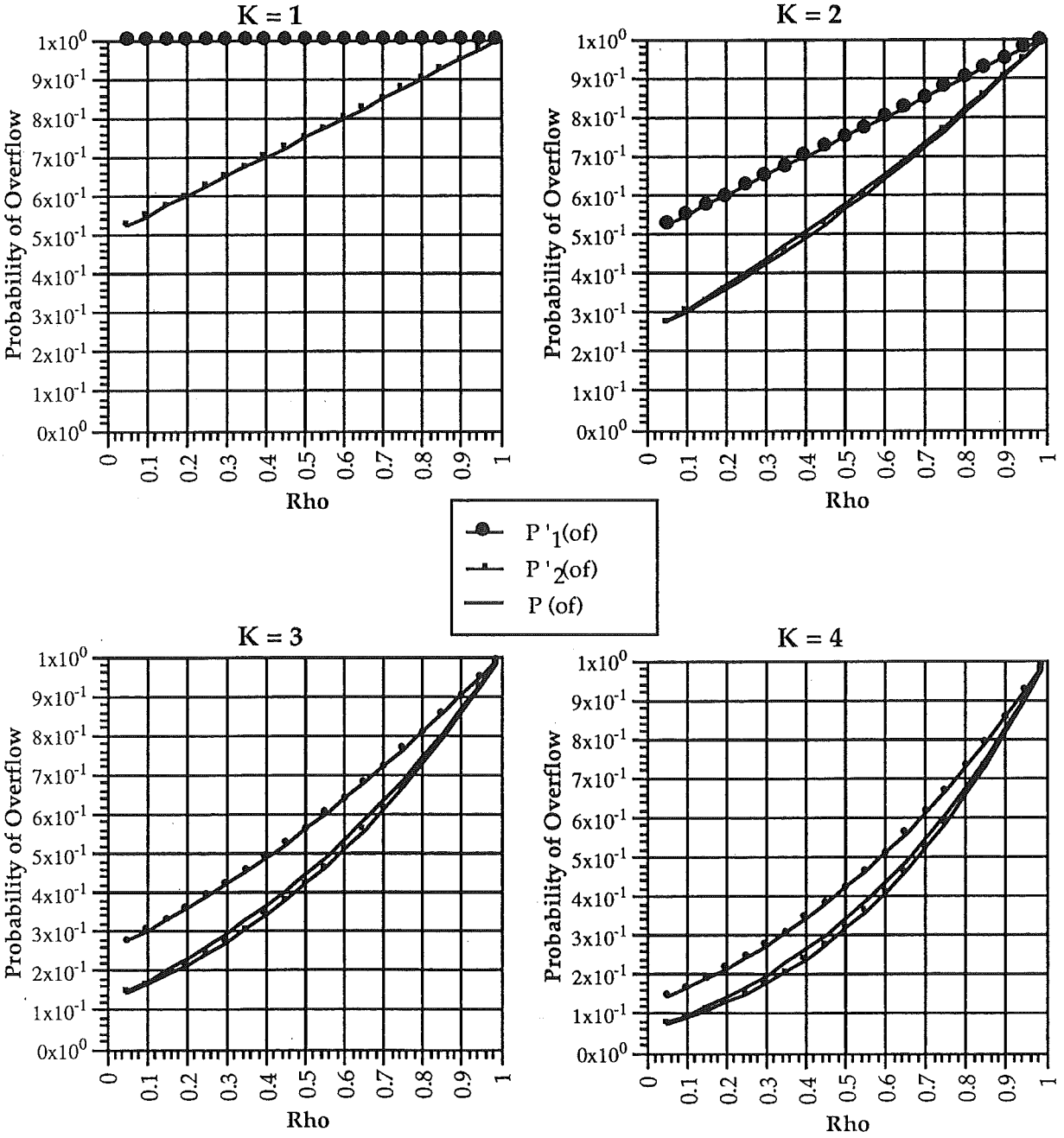
Graph 4.9

Probability of Overflow in M(b)/ M/ 1/ K
Comparison Between
P(overflow), P1' (overflow), P2'(overflow)
q = 0. 2



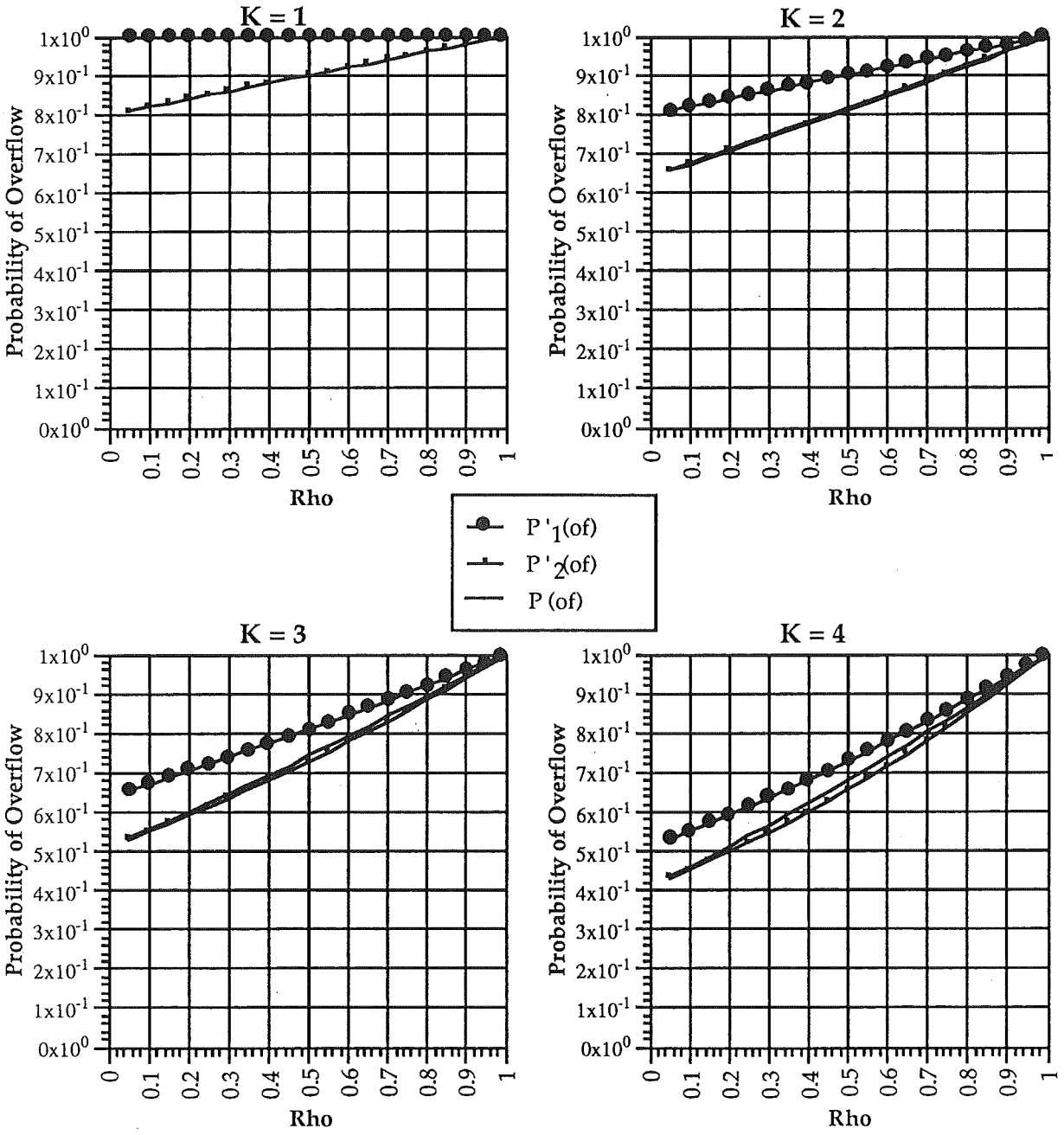
Graph 4.10

Probability of Overflow in $M^{(b)}/M/1/K$
Comparison Between
P(overflow), P_1' (overflow), P_2'(overflow)
q = 0. 5



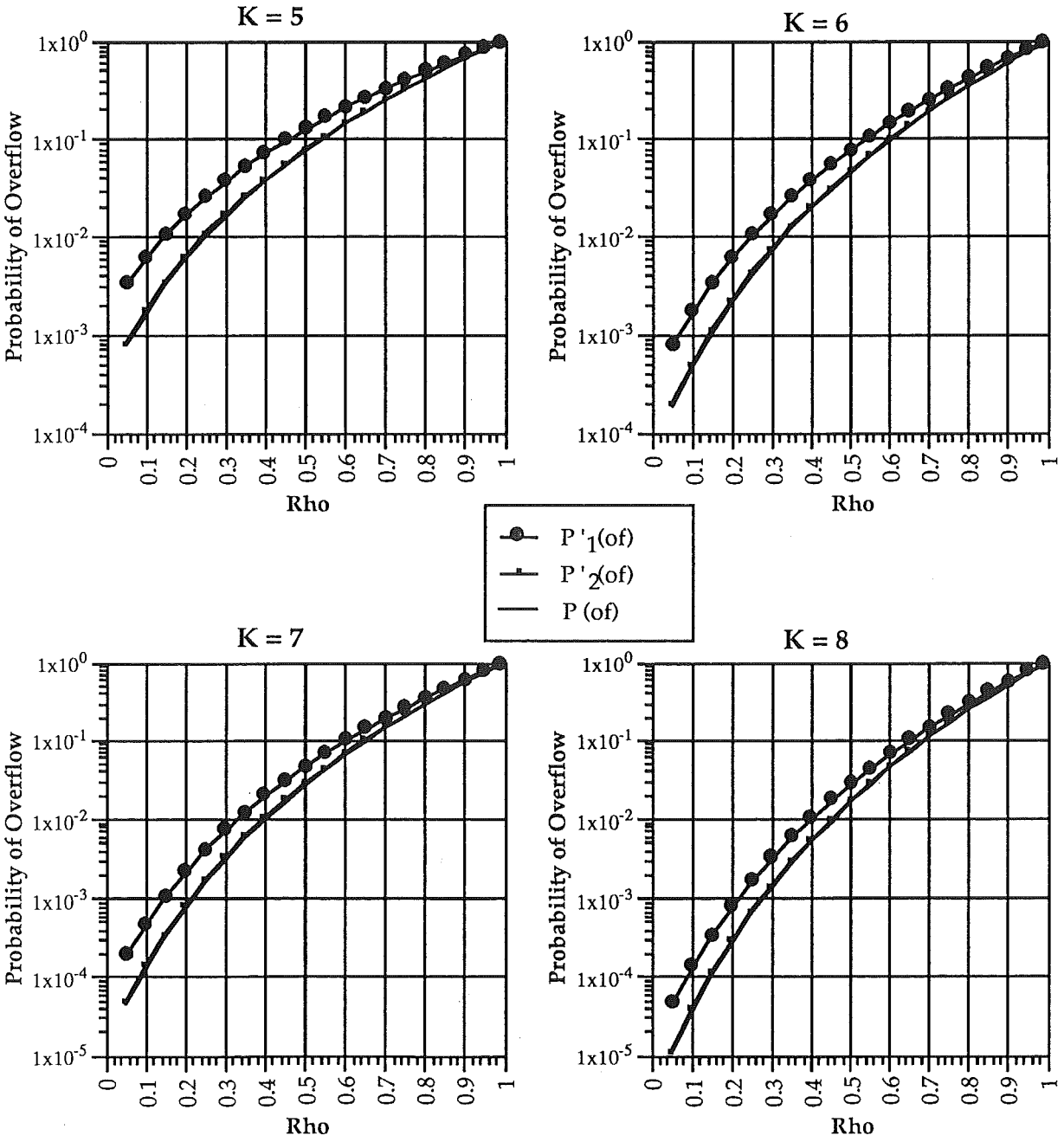
Graph 4.11

Probability of Overflow in $M^{(b)}/M/1/K$
Comparison Between
P(overflow), P_1' (overflow), P_2'(overflow)
$q = 0.8$



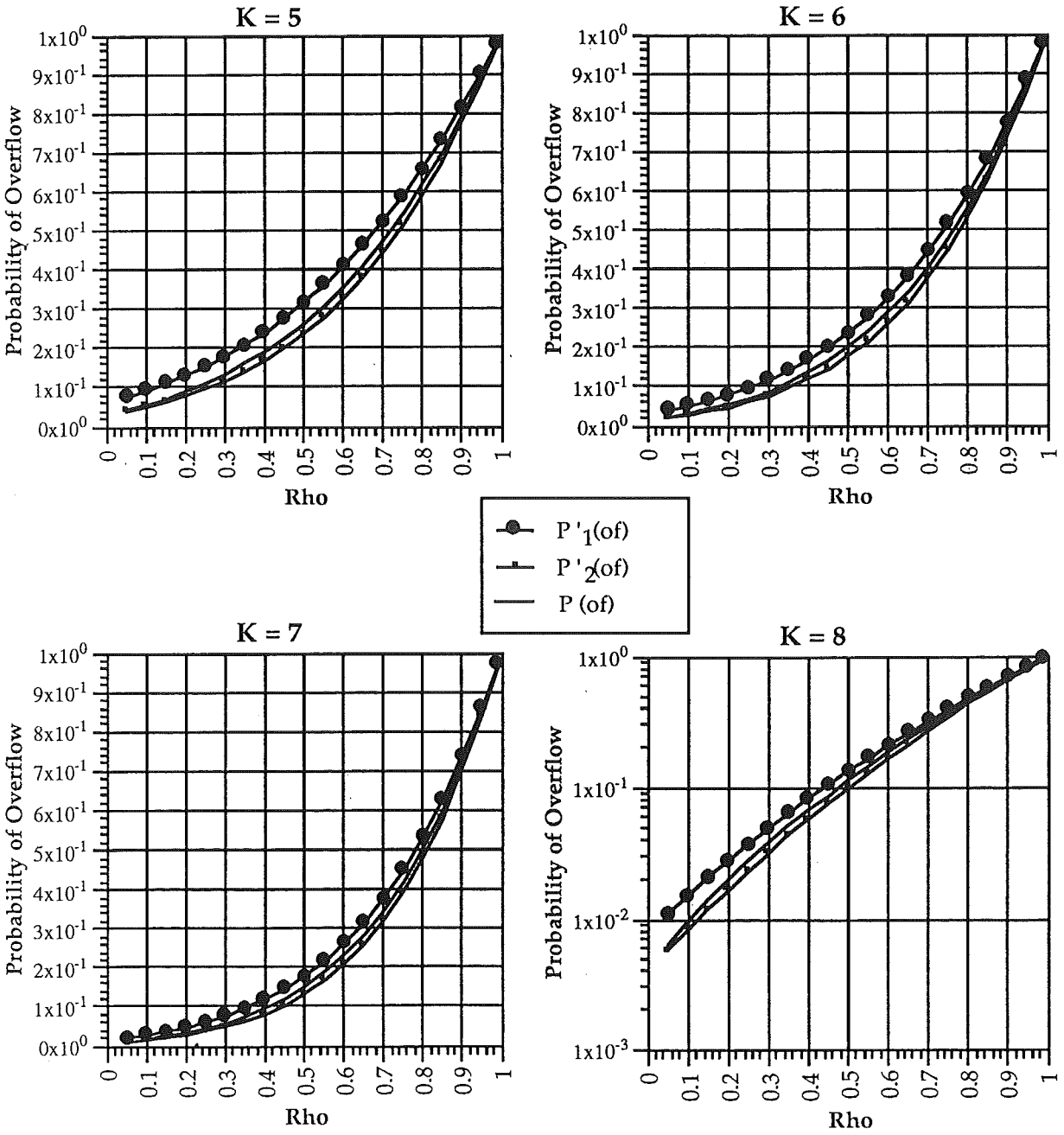
Graph 4.12

Probability of Overflow in $M^{(b)}/M/1/K$
Comparison Between
P(overflow), P_1' (overflow), P_2'(overflow)
q = 0. 2



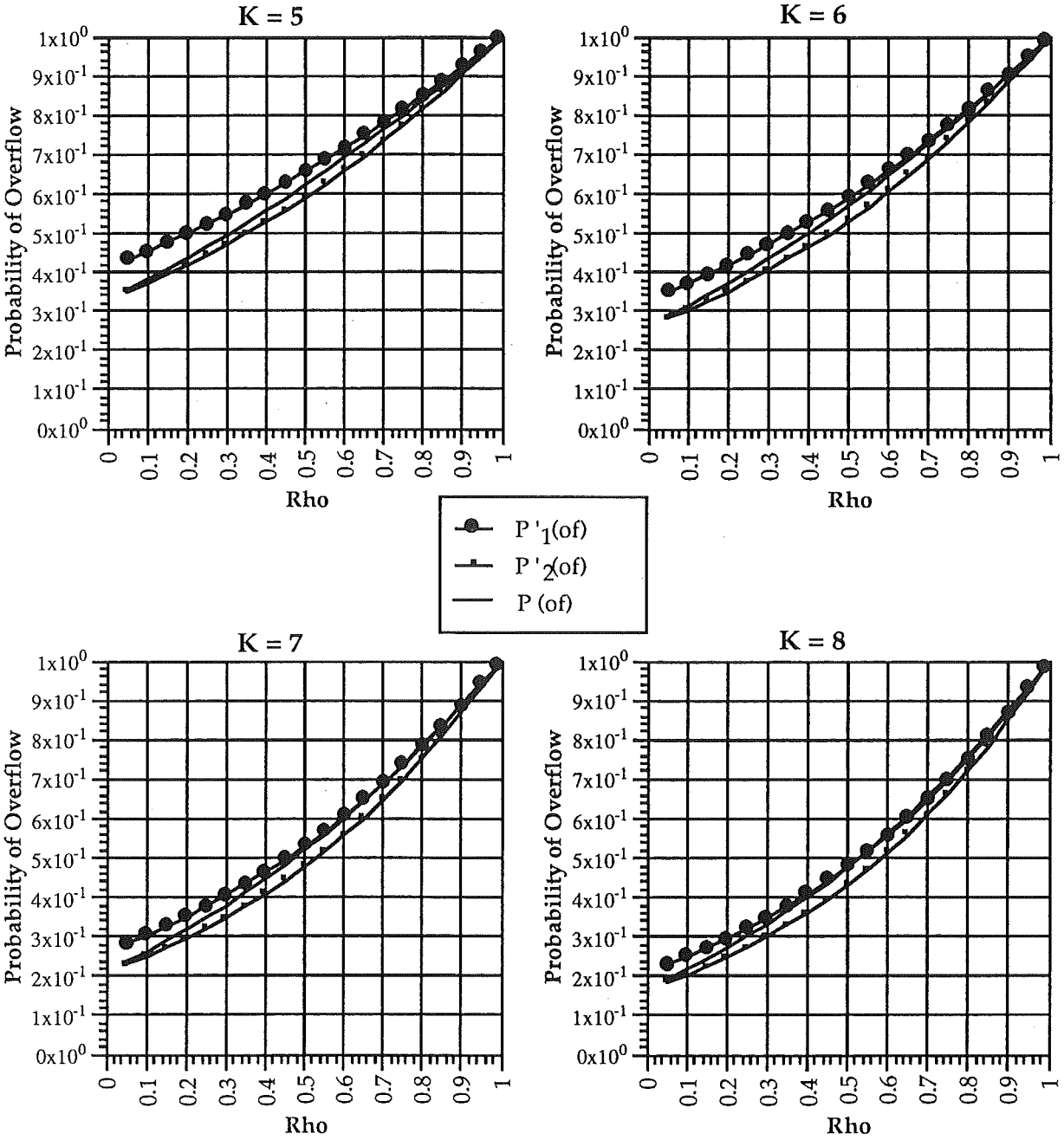
Graph 4.13

Probability of Overflow in $M^{(b)}/M/1/K$
Comparison Between
P(overflow), P_1' (overflow), P_2'(overflow)
q = 0. 5



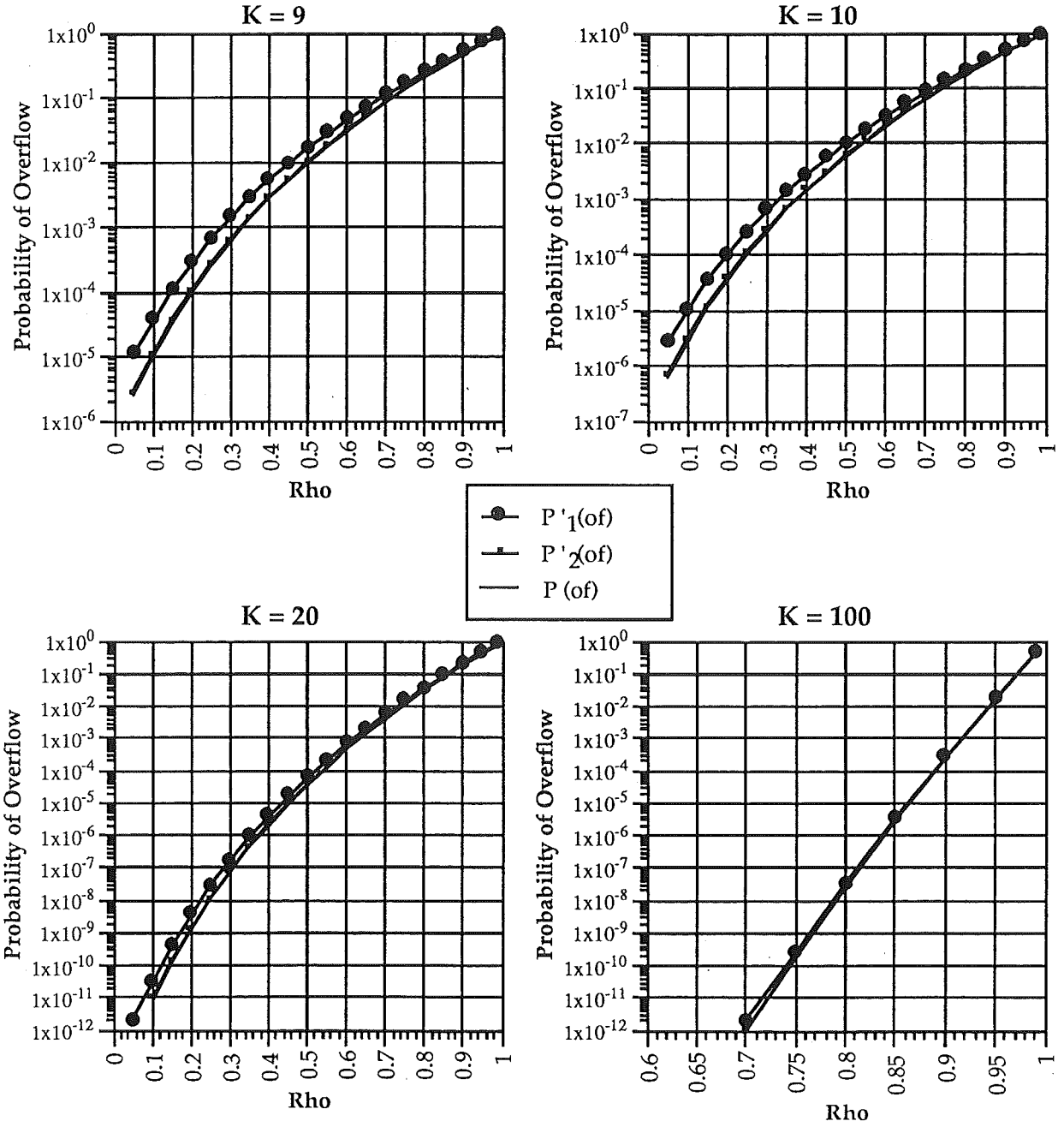
Graph 4. 14

Probability of Overflow in $M^{(b)}/M/1/K$
Comparison Between
P(overflow), P_1' (overflow), P_2'(overflow)
q = 0. 8



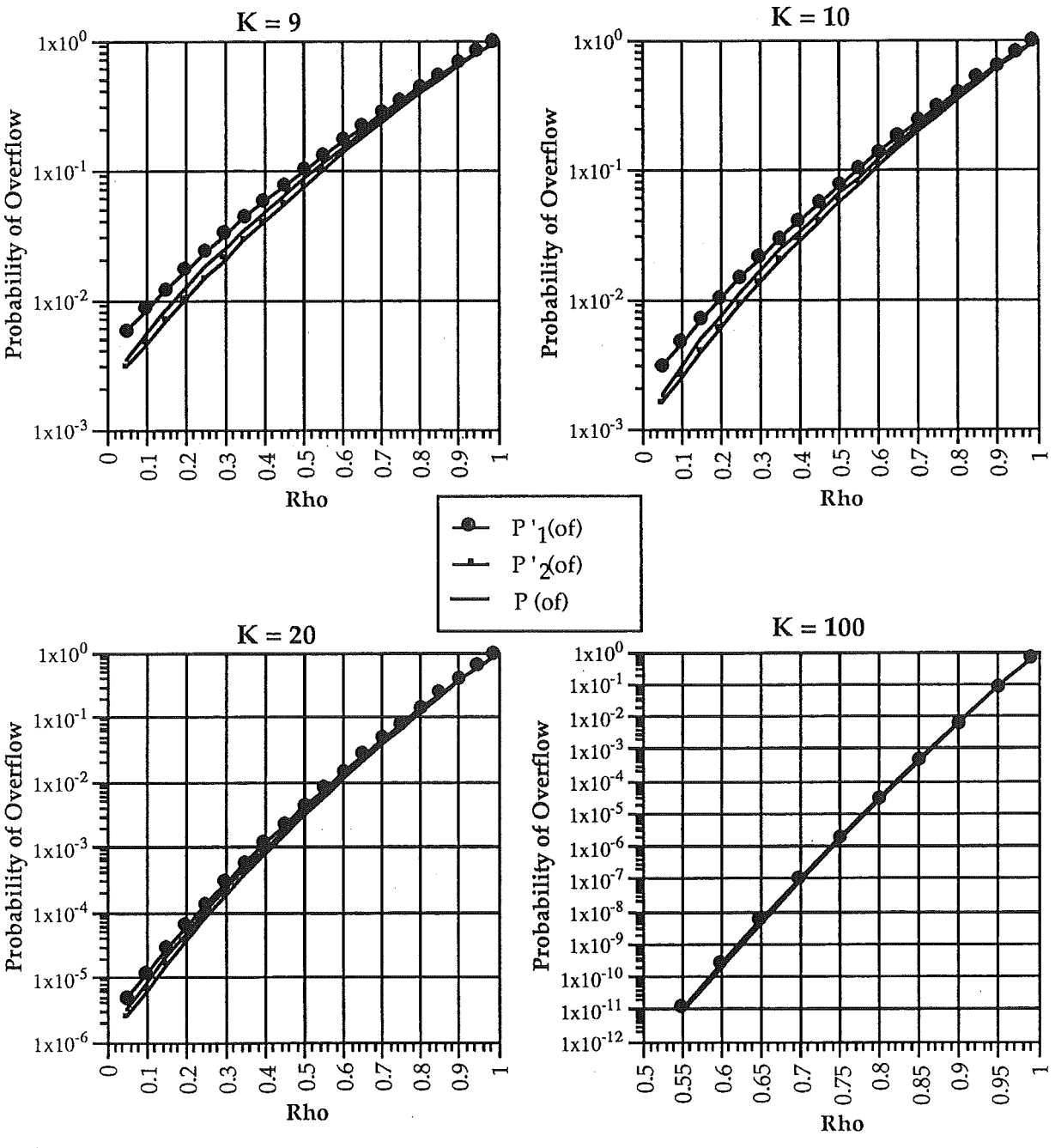
Graph 4.15

Probability of Overflow in $M^{(b)}/M/1/K$
Comparison Between
P(overflow), P_1' (overflow), P_2'(overflow)
q = 0. 2



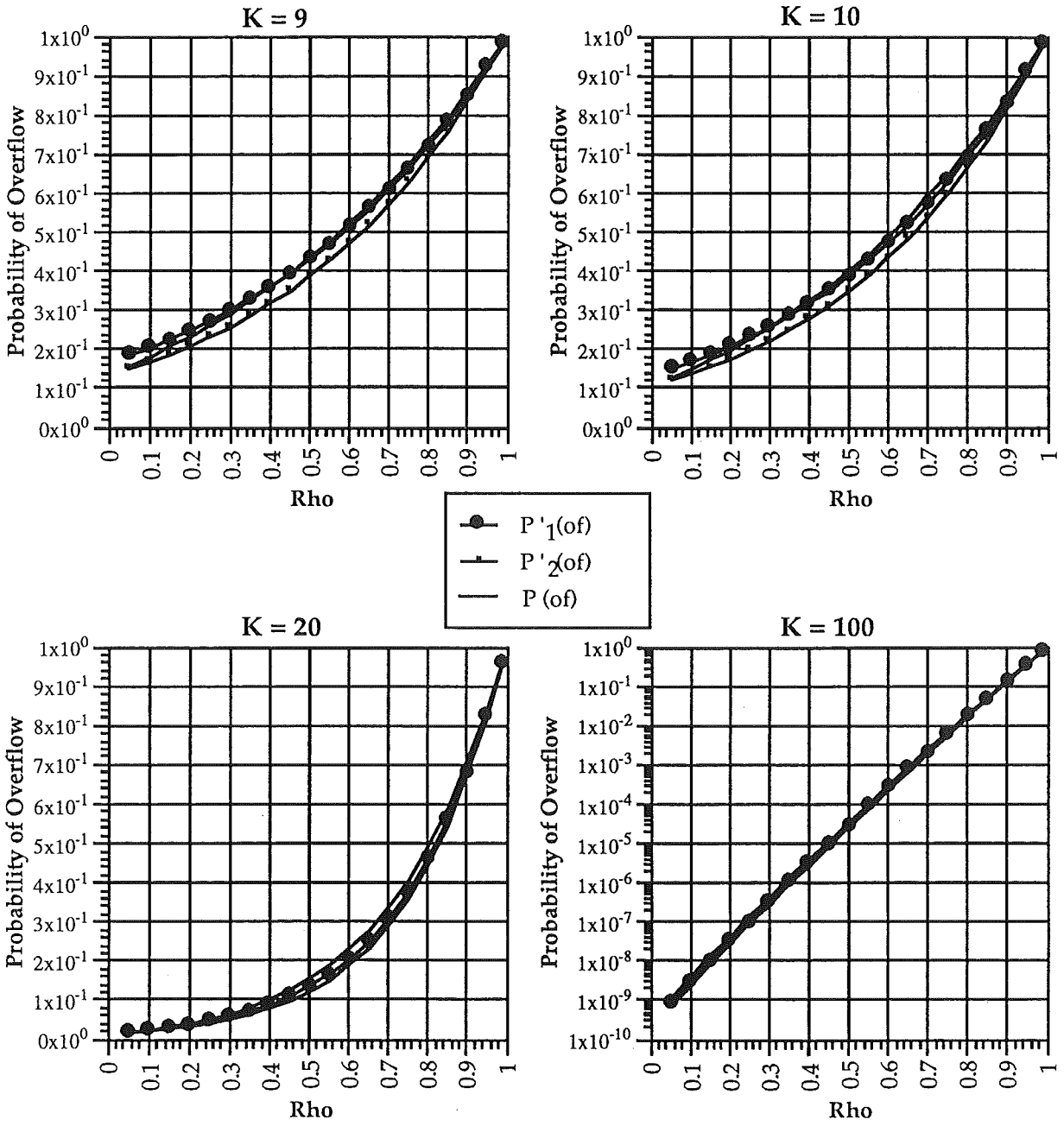
Graph 4.16

Probability of Overflow in $M^{(b)}/M/1/K$
Comparison Between
P(overflow), P_1' (overflow), P_2'(overflow)
q = 0. 5



Graph 4. 17

Probability of Overflow in M(b)/ M/ 1/ K
Comparison Between
P(overflow), P_1' (overflow), P_2'(overflow)
q = 0. 8



Graph 4.18

Chapter Five

General Conclusions and Discussion

So far, we have analysed the approximations of queueing systems of finite capacity by a system with infinite capacity in case of both individual and group arrivals. In spite of the different approaches that are used in analysis of such systems, approximations show similar property. Generally, the bigger the finite capacity is, the better the approximation.
In case of the average system waiting time approximation, the above fact is true, for all the queueing systems compared, the average system waiting time approximations are better when the system capacity is large (> 20).

Similarly, for the approximations probability of overflow approximation, the queueing system with infinite buffer overestimates the exact value of the overflow probabilities, thus they can be regarded as the 'worst-case' results for the buffer with finite capacity .

The two facts mentioned above, suggest us that it is possible to approximate performance measures of queueing system with finite buffer capacity by its infinite buffer capacity counterpart, with generally not too high error.

Appendix A

Average waiting time in system (for infinite system capacity):

$$E[W] = \frac{\sum_{i=0}^{\infty} i\, P_i}{\lambda} \qquad \text{(A-1)}$$

Average waiting time in the queue (for infinite system capacity):

$$E[W_q] = \frac{\sum_{i=0}^{\infty} (i-1)\, P_i}{\lambda} \qquad \text{(A-2)}$$

The probability of overflow for $M^{(b)}/D/1/K$ queueing system :

$$P(\text{overflow}) = P[N_K = 0]\,(1 - \sum_{i=0}^{K+1} a_i) + P[N_K = 1]\,(1 - \sum_{i=0}^{K} a_i) + \ldots\ldots \qquad \text{(A-3.1)}$$

and

$$P'_1(\text{overflow}) = P[N_\infty = 0]\,(1 - \sum_{i=0}^{K+1} a_i) + P[N_\infty = 1]\,(1 - \sum_{i=0}^{K} a_i) + \ldots \qquad \text{(A-3.2)}$$

and

$$P'_2(\text{overflow}) = P[N_\infty = 0]\,(1 - \sum_{i=0}^{K} a_i) + P[N_\infty = 1]\,(1 - \sum_{i=0}^{K-1} a_i) + \ldots \qquad \text{(A-3.3)}$$

References

[1] David G. Kendall, Some problems in the theory of queues, J. Roy. Statis. Soc Ser. B, 13, (1951), 151-173.

[2] Donald Gross and Carl M. Harris, Fundamentals of queueing theory, J Wiley & Sons, 1985, pp 156-163, 252-294.

[3] John D. C. Little, A proof of the queueing formula; L = l W, Opns. Res 9(3), (1961), 383-387.

[4] Arnold O. Allen, Probability, statistics, and Queueing theory with computer science applications, Academic Press Inc, 1990, pp 262-326.